Psychologie

Computer-unterstütztes Testen

Inaugural - Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät der
Westfälischen Wilhelms-Universität
zu Münster (Westf.)

vorgelegt von

Ulf Plessen

aus Berlin

1976

1977

Psychologie

Computer-unterstütztes Testen

Inaugural - Dissertation
zur Erlangung des Doktorgrades
der Philosophischen Fakultät der
Westfälischen Wilhelms-Universität
zu Münster (Westf.)

vorgelegt von

Ulf Plessen

aus Berlin

1976

UNIV.
LEIDEN
BIBL.

Tag der mündlichen Prüfung: 11. 2. 1977

Dekan: Professor Dr. L. Budde
Referent: Professor Dr. L. Kemmler
Korreferent: Professor Dr. H. Bommert

Für ihre verständnisvolle Hilfe und fördernde Kritik möchte ich an dieser Stelle Frau Prof. Dr. Lilly Kemmler und Herrn Prof. Dr. Hanko Bommert ganz herzlich danken; sie haben in allen Stadien der Arbeit wesentlich zu ihrem Gelingen beigetragen.

Ein besonderer Dank gilt Herrn Dieter Langheinrich, dem Leiter der Rechenstelle am Psychologischen Institut der Universität Münster. Ohne seine vielfältigen Kenntnisse und Erfahrungen bei der Programmierung und im praktischen Umgang mit Elektronischen Datenverarbeitungsanlagen, seine ständige Einsatz- und Hilfsbereitschaft wäre die Durchführung dieser Untersuchung kaum möglich gewesen.

Zu danken habe ich auch all den Studenten, die sich für die verschiedenen Untersuchungen als Versuchspersonen zur Verfügung gestellt haben.

Auch möchte ich an dieser Stelle mit Dank darauf verweisen, dass ich für zwei Jahre ein Promotionsstipendium nach dem Graduierten Förderungsgesetz erhalten habe, wodurch mir die ausschliessliche Beschäftigung mit dem Inhalt dieser Arbeit ermöglicht wurde.

GLIEDERUNG DER ARBEIT

EINFUHRUNG IN DIE FRAGESTELLUNG 1

ZUM STAND DER AUTOMATISIERUNG IN DER PSYCHOLOGISCHEN DIAGNOSTIK 4

I. Funktionen des Computers in der Psychologie 5
II. Bereiche des Computereinsatzes in der Psychologie 11
1. Uberblick über verschiedene Einsatzbereiche 11
A. Wahrnehmungspsychologie 11
B. Pädagogische Psychologie 14
C. Klinische Psychologie 18
2. Die Rolle des Computers in der Diagnostik 21
A. Die Automatisierung einzelner Phasen des diagnostischen Prozesses 22
a. Testdurchführung 22
b. Numerische Testauswertung 25
c. Qualitative Testinterpretation 26
d. Beratung 29
B. Der diagnostische Urteilsprozess 30
a. 'Black-box'-Methoden 31
b. Protokoll-Analysen 33
c. Ein normatives Modell psychologischen Urteilens 33
3. Zusammenfassung und Folgerungen 35

DER DIAGNOSTISCHE GEWINN AUFGRUND DER AUTOMATISIERTEN DURCHFÜHRUNG VON TESTVERFAHREN 38

I. Das übergreifende Untersuchungsziel 39
II. Entwicklung der technischen und programmtechnischen Grundlagen eines Systems zur Automatisierten Testdurchführung 42

1. "Hardware" 42

2. "Software" 46

A. Kurzbeschreibung der benötigten Unterprgramme 47

B. Textausgabe und Antworteingabe (Unterprogramm DAEUP) 50

C. Die programmierte Einführung in die Bedienung der Eingabetastatur (Unterprogramm TASUP) 53

III. Notwendige Voraussetzungen für die Übereinstimmung von Testergebnissen aus herkömmlicher und automatisierter Testdurchführung (1. Versuchsserie) 59

1. Entwicklung der Untersuchungsfrage 59

A. Herkömmliche und automatische Testdurchführung 59

a. Herkömmliche Testdurchführung 60

b. Verschiedene Möglichkeiten der automatischen Testdarbietung 61

aa. Das Konzept einer einfachen automatisierten Testdarbietung 61

bb. Automatisierte Testdarbietung mit Eingabekorrektur 61

cc. Einzel- und Gesamtausgabe der Testaufgaben 63

dd. Zusätzliche Faktoren 63

B. Ziel der 1. Versuchsserie 65

C. Versuchsaufbau und Versuchsdurchführung 67

2. Versuchsergebnisse 71

A. Übereinstimmung zwischen Testergebnissen aus herkömmlicher und automatisierter Testdarbietung 72

B. Auswirkungen der beiden Testsituationen 76

3. Zusammenfassung und Folgerungen 80

IV. Die diagnostische Bedeutung bei der auomatisierten Testdurchführung anfallender Informationen (2. Versuchsserie) 82

1. Herleitung der Untersuchungsfrage 83

A. Die zusätzlichen Informationsquellen 84

a. Latenzzeit 84

b. Antwortsicherheit 85

B. Ziel der 2. Versuchsserie 86
C. Versuchsaufbau und Versuchsdurchführung 88
2. Versuchsergebnisse 92
A. Aufgabenschwierigkeit 93
a. Versuchsgruppe (1) 93
b. Versuchsgruppe (2) 99
c. Vergleich der beiden Versuchsgruppen 105
B. Trennschärfe 108
a. Versuchsgruppe (1) 110
b. Versuchsgruppe (2) 115
c. Vergleich zwischen den beiden Versuchsgruppen 119
C. Innere Konsistenz 123
a. Versuchsgruppe (1) 125
b. Versuchsgruppe (2) 125
c. Vergleich der beiden Versuchsgruppen 126
3. Diskussion der Befunde 128

NEUE MÖGLICHKEITEN BEI COMPUTERUNTERSTUTZTEM TESTEN 132

I. Die bisher erarbeiteten Grundlagen 134
1. Das logistische Messmodell 134
2. tailored testing 139
A. Strategien zur Itemauswahl 140
B. Vor- und Nachteile des Verfahrens 142
3. Aufgabenorientierte Schwierigkeitsschätzung 147
A. Erste Ansätze im Bereich der CAI 148
B. Aufgabenkonstruktion nach dem logistischen Modell 152
II. Computer-unterstütztes Testen 156
1. Ein neuer Ansatz 156
2. Die Bedeutung des neuen Ansatzes für die Diagnostik 159
3. Offene Fragen 161
Zusammenfassung 162

LITERATURVERZEICHNIS 169

TABELLENANHANG 179

PROGRAMMANHANG 198

Digitized by the Internet Archive
in 2026

https://archive.org/details/IA41547208_0023

EINFÜHRUNG IN DIE FRAGESTELLUNG

Bislang besteht für den weiten Bereich einer automarisierten Diagnostik noch kein allgemein verbreiteter Oberbegriff; so ist es nicht möglich, den Gegenstandsbereich dieser Untersuchung mit einem Schlagwort zu verdeutlichen. In Analogie zur "computer-unterstützten Unterweisung" (computer-assisted instruction, CAI) könnte man von "computer-unterstützter Diagnostik" oder enger gefasst von "computer-unterstützter Testdiagnostik" (computer-assisted testing, CAT) - vgl. BÜRLI, 1975 - sprechen.

Aus der Tatsache, dass dieser Begriff noch relativ ungebräuchlich ist - er wird z.B. nicht als Stichwort in den PSYCHOLOGICAL ABSTRACTS aufgeführt - kann nun nicht unbedingt geschlossen werden, dass auf diesem Gebiet keine Forschung betrieben wird. Wir können vielmehr mit BÜRLI (1973) davon ausgehen, dass wahrscheinlich zahlreiche Untersuchungen im Bereich des CAT durchgeführt worden sind, dass aber die Berichte darüber die üblichen Informationskanäle bisher kaum erreicht haben.

Während sich für den Bereich der CAI inzwischen eine eigenständige Forschung entwickelt hat, ist für die computer-unterstützte Diagnostik noch kein integrierender Ansatz zu erkennen. Dies ist um so verwunderlicher, als zum einen die für die CAI entwickelten Computersysteme leicht auch für eine automatisierte Diagnostik nutzbar gemacht werden könnten, vor allem aber, weil man mit

HOLTZMAN (1970) computer-unterstütztes Unterrichten, Testen und Beraten als ein integriertes System auffassen sollte.

Wenn auch zur automatisierten Testverrechnung und Testinterpretation von den verschiedensten Seiten Berichte über praktische Anwendungen vorgelegt werden, so werden die Möglichkeiten einer automatisierten Testdurchführung bei weitem nicht genutzt. Zusätzlich scheinen die Arbeiten zur automatisierten Testdurchführung relativ isoliert neben denen zur Testinterpretation zu stehen (BÜRLI, 1975).

Vorliegende Arbeit beschäftigt sich im Rahmen der computerunterstützten Testdiagnostik (CAT) mit dem engeren Gebiet der automatisierten Testdurchführung. Auf diesem Gebiet beschränken sich die bisherigen Untersuchungen auf die Übernahme bereits bestehender Verfahren. Die mit der Automatisierung zusätzlich anfallende Möglichkeit, über die Zahl der richtigen Antworten hinaus weitere Informationen über den Probanden zu erhalten, wird zwar in den einschlägigen Untersuchungen erwähnt und als vorteilhaft erörtert (z.B. ELWOOD, 1969; GREEN, 1970), aber bis auf wenige Ausnahmen (z.B. DUNN, LUSHENE & O'NEILL, 1972; TOBIAS, HEDL & TOWLE, 1974) kaum praktisch genutzt.

Global lässt sich das Ziel dieser Untersuchung dahingehend bestimmen, dass die Überlegenheit der automatisierten gegenüber der herkömmlichen Testdurchführung nachzuweisen ist.

Dieser Nachweis soll von zwei unterschiedlichen Ansätzen ausgehend erbracht werden:

Zum einen wollen wir für die in der Praxis üblichen Testverfahren, die nach den Prinzipien der klassischen Testtheorie entwickelt sind, aufzeigen, dass die automatisierte Testdurchführung der herkömmlichen Darbietung zumindest dann überlegen ist, wenn wenigstens einzelne der dabei zusätzlich zu erreichenden Informationen aufgegriffen werden.

Dabei geht es nicht so sehr um die Verifizierung oder Falsifizierung einzelner, eng umschriebener Hypothesen, sondern um den Versuch, den Wert der Automatisierung für die praktische Diagnostik deutlich zu machen.

Andererseits ist zu zeigen, dass die Möglichkeiten einer automatisierten Testdurchführung bei weitem nicht genutzt werden, wenn man dem Computer nur die Aufgabe überträgt, ein bestimmtes, vorher festgelegtes Testverfahren automatisch durchzuführen. Vielmehr kann die Effektivität diagnostischen Vorgehens wesentlich dadurch erhöht werden, dass jeder einzelne Proband mit einem ihm von der Schwierigkeit der Aufgaben her genau entsprechendem Testverfahren untersucht wird. Diese unter dem Begriff "tailored testing" bekannte Vorgehensweise (z.B. LORD, 1970, 1971) scheitert in der Praxis bisher an der Grösse des dafür erforderlichen "Itempools". Ausgehend von bereits vorliegenden Untersuchungen wollen wir nachweisen, dass zumindest bei der Messung einzelner, präzise zu umschreibender Fähigkeiten der Computer auch dazu eingesetzt werden kann, im Augenblick der Testdurchführung automatisch Items von vorgegebenem Schwierigkeitsgrad zu konstruieren und diese nach bestimmten Algorithmen dem Probanden vorzugeben.

Insgesamt geht es uns hier also nicht darum, den verschiedenen Einzelansätzen für eine Automatisierung der Testdiagnostik einen weiteren, speziellen Aspekt hinzuzufügen, sondern wir wollen einen Anstoss dazu geben, nun endlich automatisierte Systeme in die diagnostische Praxis einzubeziehen.

ZUM STAND DER AUTOMATISIERUNG IN DER PSYCHOLOGISCHEN DIAGNOSTIK

Bei einem ersten Überblick über die Entwicklung der Forschung im Bereich der Psychologischen Diagnostik - und hier vor allem über die Möglichkeiten der Elektronischen Datenverarbeitung - lässt sich festhalten: Die Begeisterung über die Anwendungsmöglichkeiten des Computers, die etwa Mitte der sechziger Jahre ihren Höhepunkt erreichte (z.B. KLEINMUNTZ, 1969a), ist zum grössten Teil einer Ernüchterung aufgrund von Fragen nach der praktischen Anwendbarkeit, nach dem Verhältnis von Aufwand, Kosten und Nutzen (z.B. ELWOOD, 1972c; ROBINSON, 1973; BÜRLI, 1975) gewichen. Diese Begeisterung wurde zunächst sicher durch den Gedanken ausgelöst, dass im Prinzip jedes eindeutig zu beschreibende Problem auch durch einen logischen Automaten zu lösen ist. Gemessen an diesem Anspruch mögen die bis heute erarbeiteten praktischen Lösungen recht unbefriedigend wirken.

Diese Behauptung lässt sich sicher nicht in einem Satz beweisen oder widerlegen. Wir wollen uns daher zunächst einen gewissen Überblick über die Einsatzmöglichkeiten des Computers in der Psychologie und hier vor allem in der Diagnostik verschaffen, um auf diesem Hintergrund den Wert des bisher Erreichten besser einschätzen zu können.

Ein solcher Überblick ist unter zwei Gesichtspunkten möglich; wir können entweder von den verschiedenen Funktionen des Computers ausgehen oder von den Bereichen der Psychologie, in denen er hauptsächlich eingesetzt wird. Aus der Zielsetzung dieser Arbeit erscheint es günstiger, die funktionale Seite nur kurz anzusprechen und der anwendungsorientierten Darstellung ein grösseres Gewicht zu geben.

I. FUNKTIONEN DES COMPUTERS IN DER PSYCHOLOGIE

Mit APTER & WESTBY (1973) können wir drei Nutzungsmöglichkeiten des Computers in der Psychologie unterscheiden:

- in der Verarbeitung von Daten;
- im Bereich des psychologischen Experiments;
- in der Untersuchung von Verhaltensmodellen.

Obwohl der Bereich der Datenverarbeitung von Psychologen zur Zeit sicher am stärksten genutzt wird, soll hier auf die Funktion des Computers als Rechenanlage im engeren Sinn nicht eingegangen werden. Zwar sind eine Vielzahl von Forschungsergebnissen erst durch den Einsatz leistungsfähiger Grossrechner ermöglicht worden - man denke hier nur an die Vielzahl der berechtigten und unberechtigten Anwendungen der Faktorenanalyse, doch würde uns eine Aufbereitung dieses Einsatzbereiches zu weit von unserem eigentlichem Thema wegführen. - Es muss hier auf die Standardwerke der Statistik und auf die Programmbibliotheken der verschiedenen Hersteller von EDV-Anlagen verwiesen werden.

Bei einer Automatisierung des psychologischen Experiments ergeben sich erste Vorteile schon in der Phase der Versuchsplanung. Die zufallsmässige Zuweisung von Probanden zu einzelnen Versuchsblöcken oder die Zufallsdarbietung

verschiedener Stimuli wird durch die Benutzung von "Zufallsgeneratoren" wesentlich vereinfacht; es handelt sich dabei - vereinfacht gesagt - um Programme, durch die auf dem Computer das Ziehen zufälliger Zahlen aus einer "Lostrommel" nachgebildet werden kann.

Entscheidende Fortschritte ergeben sich durch den Einsatz des Computers bei der Versuchsdurchführung. Hier sind vor allem die Standardisierung des experimentellen Vorgehens und die präzise Erfassung der anfallenden Daten zu nennen. Auch sind bei der Übertragung der im Versuch ermittelten Daten auf verschiedene Datenträger (z.B. Lochkarten) im Rahmen der geringen Fehlerquoten der mechanischen Peripherieeinheiten des Computers (z.B. Lochkartenstanzer) weitgehend ausgeschlossen. Daneben ergibt sich für den Versuchsleiter die Möglichkeit, sofort nach Beendigung des Versuches die Ergebnisse in geordneter Form (z.B. Häufigkeitsverteilungen) ausgedruckt zu erhalten.

Der Computereinsatz bei der Untersuchung von Verhaltensmodellen wird meist unter dem Begriff "Simulation" zusammengefasst. Obwohl dieses Schlagwort, wie APTER (1973) darstellt, mit einer Reihe nicht kontrollierbarer Assoziationen belastet ist, ist es inzwischen zu einem nicht genau definierten Oberbegriff für eine Reihe von Forschungsansätzen geworden.

Gehen wir mit SEIDENSTÜCKER (1971) vom Ziel der Computersimulation aus, so können wir bei der Bildung und Überprüfung psychologischer Theorien drei sich teilweise überschneidende Anwendungsbereiche unterscheiden:

(1) Computerprogramme können zur Darstellung bereits existierender psychologischer Theorien benutzt werden; denn Computersprachen sind relativ eindeutig und stehen dabei der Struktur der psychologischen Fachsprache näher als die Formelsprache der Mathematik. Computer-

sprachen ermöglichen wie die mathematische Sprache - zum Teil jedoch mit erheblichem Programmieraufwand - die Darstellung von komplexen Zusammenhängen. Unter dieser Sichtweise von "Simulation" kann man das Programm als ein Hilfsmittel auffassen, das es dem Experimentator ermöglicht, theoretische Annahmen einfacher und genauer auf eine experimentelle Aufgabenstellung anzuwenden und dadurch aus der Theorie abgeleitete Vorhersagen überprüfen zu können.

(2) Simulationsprogramme ermöglichen die Überprüfung von Konsistenz und Vorhersagekraft psychologischer Theorien. Überspitzt könnte man fordern, dass eine verbal formulierte Theorie nur dann ausreichende Leistungsstärke besitzt, wenn alle unter dieser Theorie bisher gesammelten Beobachtungen nach Übertragung der Theorie in ein Simulationsprogramm unter den entsprechenden Versuchsbedingungen als Vorhersage oder Ergebnis erzeugt werden können. Ist ein der Theorie entsprechendes Programm erstellt worden, so lassen sich durch Veränderung einzelner Programmteile bzw. Parameterwerte Annäherung oder Abweichung der durch das Programm erzeugten Daten an die in der Realität festgestellten kontrollieren. Damit kann das Programm so entwickelt werden, dass eine möglichst gute Übereinstimmung zwischen realen und simulierten Daten erzeugt wird.

(3) Computerprogramme können bei der Erstellung psychologischer Theorien als wichtiges Hilfsmittel eingesetzt werden. Verschiedene Prozesstheorien für kognitive Funktionen wurden aufgrund der in der Denkpsychologie traditionsreichen Technik der Aufgabenprotokollanalyse und der Übertragung in Simulationsprogramme entwickelt - z.B. FELDMAN (1959), NEWELL & SIMON (1961), LAUGHERRY & GREGG (1962), FRIJDA (1967).

Während SEIDENSTÜCKER die verschiedenen Arten von Simulationsprogrammen nach den Gründen ihrer Erstellung unterscheidet (Darstellung, Überprüfung bzw. Entwicklung einer psychologischen Theorie), berücksichtigt APTER (1973) stärker den Ausgangspunkt ihrer Entstehung. So unterscheidet APTER zwischen Modellen, die von

- einer Theorie,
- dem beobachteten Verhalten oder
- in der Praxis erforderlich gewordenen Computerprogrammen

abgeleitet werden.

Vergleicht man diese beiden Einteilungsschemata, so entspricht APTERs Kategorie der aus einer Theorie abgeleiteten Simulationmodelle etwa den Kategorien Darstellung und Überprüfung einer psychologischen Theorie bei SEIDENSTÜCKER. Auch stimmt in etwa APTERs Kategorie der aus dem beobachteten Verhalten abgeleiteten Programme mit SEIDENSTÜCKERS Zusammenfassung von Simulationsprogrammen als Hilfsmittel bei der Erstellung einer psychologischen Theorie überein.

Ein grosser Teil von Simulationsprogrammen, die durch SEIDENSTÜCKERs mehr theoretische Einteilung nicht erfasst werden können, berücksichtigt APTER in der Kategorie jener Programme, die ausgehend von zunächst in der Praxis erforderlichen Problemlösungsprogrammen dann auch als Simulationsprogramme aufgefasst werden können. Damit berücksichtigt APTER wesentlich stärker die praktischen Gegebenheiten, und wir wollen deshalb auf seine dritte Kategorie hier noch etwas näher eingehen.

Ausgangspunkt für die Erstellung von Simulationsprogrammen, die auf nach praktischen Erfordernissen entwickelten Programmen aufbauen, ist nicht der Wunsch nach Überprüfung oder Erstellung einer Theorie für bestimmte - z.B. intellektuelle Prozesse, sondern die Notwendigkeit, ein umschrie-

benes Aufgabengebiet so weit zu erarbeiten, dass danach die Lösung der Aufgabe einem Automaten übertragen werden kann. Erst in einem weiteren, aber nicht unbedingt erforderlichen Schritt können solche Programme dann als Prozessmodelle aufgefasst werden. Ein vereinfachtes Beispiel:

Die Entwicklung von Klarschriftlesern, also von Geräten, mit denen maschinen- oder auch handgeschriebene Texte ohne den Umweg der Übertragung auf einen maschinenlesbaren Datenträger (z.B. Lochkarten) in den Computer direkt zur Verarbeitung eingegeben werden können, wurde aus Rationalisierungsgründen in der kommerziellen Datenverarbeitung - man denke hier nur an die Menge der im Bankwesen zu verarbeitenden Belege - notwendig. Unabhängig von dem Entstehungsgrund können diese Programme aber auch als Simulationsmodelle für die menschliche Mustererkennung aufgefasst und zur Bildung oder Überprüfung von Theorien zur Mustererkennung herangezogen werden.

Für diese Gruppe von Simulationsmodellen sind meist "heuristische" Problemlösungsstrategien kennzeichnend. Die Programme enthalten zur Lösung eines Problems eine Reihe unterschiedlicher Strategien, von denen im konkreten Fall durch planvolles Probieren jeweils die erfolgreichste ausgewählt wird. Auch wird der Erfolg der einzelnen Strategien ständig überwacht und registriert. Damit erhöht sich mit zunehmender "Erfahrung" des Programms die Wahrscheinlichkeit, dass sofort die den grössten Erfolg versprechende Strategie eingesetzt werden kann.

Im Gegensatz zu "algorithmischem" Vorgehen garantieren heuristische Methoden jedoch nicht unbedingt den Erfolg. Wir wollen das an einem einfachen Beispiel von APTER (1973) erläutern:

Für eine grosse Bibliothek existiere nur ein nach Autorennamen, jedoch kein nach Titeln geordneter Katalog. Die Aufgabe soll darin bestehen, möglichst schnell ein Buch aufzufinden, von dem zwar der Titel, aber nicht der Autor bekannt ist. Ein Suchprogramm wird dann auf jeden Fall erfolgreich sein, wenn es nach dem Algorithmus vorgeht, die Autorenkartei solange durchzugehen, bis der gesuchte Titel gefun-

den ist. Wesentlich schneller, aber nicht unbedingt erfolgreich arbeitet die Heuristik, vom Titel her auf mögliche Autoren zu schliessen und dann unter dem entsprechenden Autorennamen zu suchen.

Heuristische Methoden müssen zur Problemlösung zumindest dann eingesetzt werden, wenn (noch) kein Algorithmus existiert. So gibt es noch keinen Algorithmus, der den Sieg im Schachspiel garantiert, jedoch sind eine Reihe von Heuristiken entwickelt worden, die die Chancen wesentlich verbessern.

Für Problemlösungsprogramme, die von heuristischen Methoden ausgehen, wird häufig das Schlagwort von der "künstlichen Intelligenz" gebraucht. Damit wird vor allem der Aspekt angesprochen, dass heuristische Modelle im Gegensatz zu Programmen, die die Problemlösung durch die Anwendung beliebig vieler Gleichungen erreichen, wesentlich lebensnaher und "intelligenter" wirken. Unberücksichtigt bleibt bei dieser Sichtweise jedoch, dass die "Intelligenz" des Programms nur die Intelligenz des Programmautors widerspiegelt.

II. BEREICHE DES COMPUTEREINSATZES IN DER PSYCHOLOGIE

Es lässt sich kaum ein Teilbereich der Psychologie finden, in dem der Computer - abgesehen von seiner Funktion als Rechenanlage im engeren Sinn - nicht erfolgreich eingesetzt werden könnte oder schon eingesetzt worden ist. Eine Bestandsaufnahme der vielfältigen Einsatzmöglichkeiten muss sich daher auf eine Reihe typischer Beispiele beschränken.

Im ersten Teil dieses Abschnitts wollen wir Untersuchungen aus den Gebieten der Wahrnehmungspsychologie, der Pädagogischen und Klinischen Psychologie aufführen. Dabei wird die Bedeutung dieser Untersuchungen für den Einsatz des Computers bei der Automatisierung der Diagnostik meist ohne ausdrücklichen Hinweis deutlich. Im zweiten Teil ist dann aufzuzeigen, wie weit die Automatisierung in der Diagnostik bisher fortgeschritten ist.

1. ÜBERBLICK ÜBER VERSCHIEDENE EINSATZBEREICHE

A. Wahrnehmungspsychologie

Vor allem für Experimente im Bereich der Psychologie der Wahrnehmung werden häufig optische Vorlagen als Reizmaterial benötigt. Statische Vorlagen in der Art von Kurven und Histogrammen lassen sich ohne grossen Programmieraufwand über den Computer erzeugen und können auf einem Kathoden-Strahl-Oszilllographen ausgegeben werden. Dazu wird das Oszillographenbild üblicherweise in 256 * 256 oder 1024 * 1024 Rasterpunkte[1)] zerlegt, die über ihre Koordinaten vom

---

1) Diese auf den ersten Blick vielleicht nicht "rund" erscheinenden Zahlen ergeben sich aus der in der EDV notwendigen Zahlendarstellung im Dualsystem (Basis 2).

Computer aus einzeln angesprochen werden können und dann aufleuchten. Die Koordinaten der einzelnen Bildpunkte können entweder vor Versuchsbeginn im Computer gespeichert werden und werden dann bei Bedarf abgerufen und ausgegeben, oder sie werden erst zum Zeitpunkt der Bildausgabe - vor allem bei durch Kurvengleichungen eindeutig zu beschreibenden Figuren errechnet. Zusätzlich lassen sich auch Zeitpunkt und Dauer der Bildausgabe genau steuern, so dass bei Beachtung der Verschmelzungsfrequenzen für die optische Wahrnehmung bewegte Bilder erzeugt werden können. Auch lässt sich damit die Bildausgabe über den Computer wie ein Tachistoskop einsetzen, d.h. statische Vorlagen können beliebig lang oder kurz dauernd ausgegeben werden. Jedoch im Gegensatz zu den mechanischen Tachistoskopen, bei denen meist nur zwei oder drei verschiedene Vorlagen abrufbar sind (zwei- bzw. drei-Kanal-Tachistoskop), ermöglicht die Computer-Ausgabe die Benutzung einer Vielzahl von Kanälen.

UTTAL (1969) benutzte diese Vielzahl von Kanälen zur Erzeugung von "tanzenden Punktmustern" im Sinne eines visuellen dynamischen Rauschens. Dabei variierte er die Anzahl der Punkte und ihre Bewegungsgeschwindigkeit als Maske über einem gleichzeitig dargebotenen Buchstaben. Wurde genügend dynamisches Rauschen zusammen mit dem Buchstaben dargeboten, war dieser nicht zu erkennen. Jedoch genügte den Versuchspersonen die alleinige Darbietung des Buchstabens für Bruchteile einer tausendstel Sekunde vor oder nach dem Einsetzen des Rauschens zur Identifikation des Buchstabens.

Neben der genau reproduzierbaren Reizdarbietung gestattet die automatisierte Versuchsdurchführung die präzise Erfassung der Versuchspersonenreaktionen und der dabei auftretenden Reaktionszeiten. Die gemessenen Daten können dann - wie üblich - nach dem Versuchsende ausgewertet werden. Darüber hinaus ist es bei solchen "on-line-Experimenten", d.h. bei Experimenten, die direkt über einen Computer gesteuert werden, möglich, den Fortgang des Versuchs aufgrund einer gleichzeitig durchgeführten Analyse der bisher angefallenen Ergebnisse von Schritt zu Schritt zu steuern.

Eine über logische Schaltkreise steuerbare Bildröhre - die Vielfalt der Schaltmöglichkeiten eines Computers war in diesem Experiment nicht erforderlich - setzte KELLEY (1969) bei seinem 'adaptive measurement system' zur Untersuchung optischer Unterschiedsschwellen ein. Unter Unterschiedsschwelle wird die Abweichung verstanden, um die sich z.B. bei Längenschätzungen zwei Strecken unterscheiden müssen, damit sie gerade noch als unterschiedlich erkannt werden. Im Gegensatz zu herkömmlichen Schwellenuntersuchungen, bei denen beim Kontanzverfahren ein nicht variierter Standardreiz mit um konstante Beträge variierten Vergleichsreizen, die jedoch in einer Zufallsreihenfolge dargeboten werden, zu vergleichen ist, arbeitete KELLEY nicht mit konstanten Reizvariationen, sondern diese wurden nach der Diskriminierungsfähigkeit der Versuchspersonen verändert. Bei richtiger - über Knopfdruck dem System mitgeteilter Antwort - wurden die Abstände zwischen den verschiedenen Vergleichsreizen kleiner, bei falscher Einschätzung wurden sie grösser.

In diesem Sinn kann der Computer zu einer genauen und "unerschütterlichen" Hilfskraft im psychologischen Labor werden. Aber wollte man ihn nur auf diese Rolle beschränken, so würde damit nur ein geringer Teil der gebotenen Möglichkeiten genutzt.

Die weitere Möglichkeiten des Computereinsatzes ausnutzenden Simulationsstudien beschäftigen sich im Bereich der Wahrnehmung vor allem mit der Mustererkennung (pattern recognition). Ein Teil dieser Entwicklung ist sicher ökonomisch zu begründen - man denke hier nur an die Entwicklung von Klarschriftlesern für den kommerziellen Bereich, wo die technischen Schwierigkeiten der Mustererkennung über komplexe Systeme zum grössten Teil gelöst ist. Doch ergaben sich damit auf der anderen Seite die auch psychologisch relevanten Fragen nach der Isolierung der "pattern" vom Hintergrund, nach ihrer Erkennung und nutzvollen Klassifizierung.

WIDROW & SMITH (1964) entwickelten mit dem Programm ADALINE ein mit zunehmender Erfahrung lernendes System zur Buchstabenerkennung. Vor den praktischen Einsatz ist eine Lehrphase geschaltet, während der das System in ständigem Dialog mit einem menschlichen Lehrer steht. Wenn das Programm

glaubt, einen Buchstaben richtig erkannt zu haben, teilt es seine Erkenntnis dem Lehrer mit, und erhält dann eine entsprechende Rückmeldung.

CASEY & NAGY (1971) beschreiben ein Programm, das es ohne Hilfe eines menschlichen Lehrers lernt, maschinengeschriebenen Text richtig zu erkennen. Im ersten Schritt werden durch iterative Clusteranalysen die eingelesenen Buchstaben in einzelne Cluster getrennt. Dabei treten Fehler bald nur noch bei sehr selten vorkommenden Buchstaben wie dem "x" auf. Damit kann das Programm zwar die eingelesenen Buchstaben als Muster voneinander trennen, doch fehlt die Zuordnung zwischen Muster und dem entsprechenden "Buchstabenwert". Dieses Problem wird in einem zweiten Schritt über einen "alten Trick" zur Entschlüsselung von Texten gelöst. Bei einem neuen Durchlauf ermittelt das Programm die Wahrscheinlichkeiten für das Auftreten der einzelnen Buchstabenmuster und vergleicht diese Wahrscheinlichkeiten mit einer vorher eingegebenen Tabelle, die die Wahrscheinlichkeiten für das Auftreten der einzelnen Buchstaben in englischsprachigen Texten enthält. Fast fehlerfrei kann das Programm maschinengeschriebenen Text dann lesen, wenn statt der Wahrscheinlichkeiten für die einzelnen Buchstaben die Wahrscheinlichkeiten für die Kombination von je zwei Buchstaben benutzt werden.

B. Pädagogische Psychologie

Zusammen mit der sprunghaften Entwicklung der Computertechnologie wurden auch die Anwendungsmöglichkeiten des Computers zu Unterrichtszwecken immer deutlicher. Einen wesentlichen Anstoss zur Entwicklung der CAI (computer-assisted instruction) gaben die Arbeiten von B.F. SKINNER, der Lehrmaschinen benutzte, die nach den Prinzipien des operanten Konditionierens arbeiteten.

Ehe wir auf die Rolle des Computers in der CAI zu sprechen kommen, sind kurz die wesentlichen Arten der programmierten Unterweisung darzustellen. Dabei ist zu beachten, dass "programmierte Unterweisung" nicht unbedingt Unterweisung durch einen (programmierbaren) Computer bedeutet.

Die linear programmierten Unterweisungen beruhen auf der Technik des "shaping". Allgemein bedeutet "shaping", dass zunächst jede Reaktion, die in Richtung auf ein angestrebtes Verhaltensmuster zeigt, verstärkt wird. Danach werden nur noch solche Verhaltensweisen verstärkt, die über das vorher gezeigte Verhalten herausgehen.

Eine linear programmierte Unterweisung übernimmt diese Technik, indem am Ende einer jeden Unterrichtseinheit (frame) aus dem bisher bearbeiteten Stoff leicht zu beantwortende Fragen gestellt werden. Die Lösungen auf diese Fragen werden zu Beginn der nächsten Unterrichtseinheit noch einmal dargeboten. Dabei wird angenommen, dass der Lernende durch die Möglichkeit, fast immer die richtige Antwort geben zu können, verstärkt wird. - Solcher Art programmierter Unterweisungen können auch in Buchform oder mit eigens entwickelten Maschinen dargeboten werden.

Die Methode des verzweigten Programms wurde vor allem von CROWDER (1960) entwickelt. Um das langsame und manchmal auch recht langweilige Vorgehen linearer Programme zu vermeiden, werden bei dieser Technik Zwischenfragen nicht so sehr zur Verstärkung sondern vor allem zur Überprüfung des bisher erworbenen Wissens eingesetzt. Hauptbestandteil verzweigter Programme sind wiederum eine Reihe aufeinander aufbauender Unterrichtseinheiten. Nach jeder Einheit wird in Form einer Mehrfach-Wahl-Antwort (multiple choice) der Lernfortschritt kontrolliert. Bei richtiger Lösung führt das Programm zur nächsten Einheit, wo die Richtigkeit der Lösung zur Verstärkung noch einmal bestätigt wird. Bei falschen Antworten wird je nach Art des Fehlers zu einer den Hauptlernschritt ergänzenden Einheit verzweigt, von wo der Weg über eine Wiederholung der zunächst falsch bearbeiteten Einheit dann zum nächsten Lernschritt führt. Natürlich ist es möglich, in die verzweigten Einheiten wiederum Kontrollfragen einzuarbeiten und ggf. noch weiter zu verzweigen. - Auch verzweigte Programme können in Buchform oder auf speziellen Maschinen dargeboten werden.

Beim Übergang von der programmierten Unterweisung zu der computer-unterstützten Unterweisung (CAI) kann der Computer leicht die Steuerungsfunktionen der Lehrmaschine übernehmen, wodurch die Verzweigungsmöglichkeiten fast unbegrenzt erweitert werden. Die Steuerung verzweigter Lernprogramme durch einen Computer stellt einen ersten Schritt zur CAI dar und wird durch entsprechende höhere Programmiersprachen - z.B. COURSEWRITER (GRUBB & SELFRIDGE, 1963) - wesentlich erleichtert.

Diese Programmiersprache stellt dem Programmierer eine Reihe von Befehlen zur Verfügung, durch die Ausführung komplexer Operationen auf dem Computer ermöglicht wird. Wird z.B. eine Lerneinheit mit dem Befehl QU ("question") gekennzeichnet, so wird damit neben der Ausgabe des entsprechenden Textes, z.B. auf einen Bildschirm, auch das Warten des Computers auf die Antwort des Schülers und das Speichern der eingegebenen Antwort bewirkt.

Bleiben die Aufgaben des Computers jedoch nur auf Steuerung und Ausgabe vorher festgelegter Programme beschränkt, so ersetzt der Computer im Grunde nur eine wesentlich einfacher angelegte Lehrmaschine. UTTAL (1967) kennzeichnet etwas überspitzt die nur auf diese Weise für die CAI eingesetzten Computer als "degenerate computer teaching machines".

Mit Hilfe der vielfältigen Möglichkeiten eines Computers kann der Bereich der programmierten Unterweisung wesentlich erweitert und flexibler gestaltet werden. Einerseits ergibt sich im Dialog mit der Maschine eine wesentliche Erleichterung durch Programme, die den Dialog nicht nur in codierter Form sondern auch in natürlicher Sprache erlauben.

Als Beispiel kann das Programm STUDENT von BOBROW (1964) angeführt werden. Student löst eingekleidete Rechenaufgaben etwa von einem Schwierigkeitsgrad, der dem der Unterstufe weiterführender Schulen entspricht. Wie für den Schüler besteht für das Programm das Problem nicht darin, die Rechnung zu bewältigen, sondern in dem Aufgabentext die eigentliche Rechenaufgabe zu erkennen. Rein quantitativ lässt sich das daran ablesen, dass auf einen Rechenbefehl etwa tausend Umformungsbefehle kommen.

STUDENT ist wie fast alle Problemlösungsprogramme - als "klassisches" Beispiel ist vor allem der GENERAL-PROBLEM-SOLVER von NEWELL, SHAW & SIMON (1957) zu nennen - ein heuristisches Programm. Eine Besonderheit dieses Programms ist die Rückfrage beim Programm-Benutzer, falls das Programm von sich aus keinen Weg zur Lösung einer Rechenaufgabe findet. STUDENT fragt dann, ob dem Benutzer Beziehungen zwischen bestimmten, in der Rechenaufgabe enthaltenen Variablen bekannt sind. Nach Eingabe solcher Beziehungen sucht das Programm dann erneut nach einem Lösungsweg.

Zum anderen eröffnet der Computereinsatz in der CAI die Möglichkeit, die Schwierigkeitsstufung des Lehrstoffes dem Lernfortschritt des Benutzers anzupassen.

Das Arithmetik-Programm COACH (WEXLER, 1970) arbeitet mit einer dynamischen Entscheidungsprozedur für verschiedene Ebenen der Aufgabenschwierigkeit. Es enthält acht Schwierigkeitsebenen, die zur feineren Abstufung jeweils in zehn Schwierigkeitsgrade unterteilt sind. Der Schwierigkeitsgrad einer neu zu bearbeitenden Aufgabe richtet sich nach dem der vorher bearbeiteten; er steigt wenn zwei von drei einander folgenden Aufgaben richtig beantwortet werden. Er sinkt, wenn zwei Aufgaben nacheinander nicht gelöst werden. Eine weitere Entscheidungshilfe, die es dem Programm gestattet, besser auf die Bedürfnisse des Einzelnen einzugehen, ist eine Art "Lernfaktor". Üblicherweise werden zu jedem Schwierigkeitsgrad jeder Ebene zwei Fragen gestellt. Beantwortet ein Schüler nun jede Frage richtig, geht das Programm in immer grösseren Schritten vorwärts.

Die von dem Programm SOCRATES (STOLUROW, 1965, 1967) benutzten Entscheidungsregeln beziehen noch weitere Informationen über den Schüler mit ein. Für jeden Lehrstoff gibt es eine Anzahl unterschiedlicher Lehrprogramme. Aufgrund eines Lehrstoff-bezogenen Vortests und einer allgemeinen Leistungsprüfung, sowie unter Berücksichtigung der Schwierigkeit des vom Schüler angestrebten Lernziels und der dafür zur Verfügung stehenden Zeit wird zunächst entschieden, welches Lehrprogramm für den einzelnen Schüler am besten geeignet ist. Während der eigentlichen Unterrichtsphase wird der Erfolg des Schülers dann nach zwei Gesichtspunkten beurteilt. Ähnlich wie bei COACH richtet sich die Schwierigkeit einer Aufgabe jeweils nach dem bisherigen Verlauf; darüber hinaus wird jedoch auch das eingesetzte Lehrprogramm ständig auf seinen didaktischen Nutzen überprüft und gegebenenfalls durch ein anderes ersetzt.

C. Klinische Psychologie

Im engeren Bereich der Klinischen Psychologie wird der Computer zur Zeit vor allem als Hilfsmittel bei der wissenschaftlichen Theorienbildung und -überprüfung eingesetzt. Dabei handelt es sich vor allem um erklärende Modelle.

MOSER, v. ZEPPELIN & SCHNEIDER (1970) entwickelten ein Modell zur Simulation neurotischer Abwehrprozesse. Das Modell ist zunächst auf die Abwehrorganisation der Ich-Struktur beschränkt; im Sinne der Psychoanalyse wird davon ausgegangen, dass die im "Ich" unbewusst erfolgenden Abwehrmechanismen sich auf solche Triebimpulse des "Es" beziehen, deren Befriedigung aufgrund der Anforderungen des "Über-Ich" und der Notwendigkeiten der Realität dem "Ich" unmöglich sind. Dabei beabsichtigen die Autoren nicht, die Herkunft neurotischer Konflikte zu erklären, sondern es werden lediglich die Abwehrprozesse untersucht, die bei einem Konflikt eingesetzt werden. Auslöser für den Konfliktfall ist ein aktueller Triebablauf. In Anlehnung an die Psychoanalyse gehen in das Modell orale, anale oder genitale Teiltriebe ein, d.h. für die Entwicklung neurotischer Abwehrprozesse spielt die meist in den entsprechenden Phasen der frühen Kindheit einsetzende Abwehr verbotener Triebregungen die entscheidende Rolle.

Mit dem Simulationsprogramm ARTIFICIAL PARANOIA versuchten COLBY, WEBER & HILF (1971) die Abläufe paranoider Informationsverarbeitung nachzubilden, also solche Formen wahnhafter Wahrnehmungen und Wahrnehmungsverarbeitungen wie sie vor allem für bestimmte Formen der Schizophrenie symptomatisch sind. Als Wahninhalte enthält das Programm die Erlebnisse physischer Bedrohung, des Erniedrigt- und Kontrolliert-Werdens. Treten in einem simulierten Gespräch bestimmte Signalbegriffe auf, die den paranoiden Komplex - besonders sensitiv sind die Bereiche um die Begriffe "Selbst" und "Selbstwert" - aktivieren können, so reagiert das Programm entsprechend der vorher festgelegten Distanz zwischen Komplex und Signalbegriff. Die Reaktionen auf bedrohliche Signale laufen über die Variablen Angst, aggressive Erregung und Misstrauen. Je nach Zustand dieser Variablen reagiert das Programm mit Angst auf die Äusserungen seines Gegenübers und bricht bei einem bestimmten Angstausmass den Kontakt ab.

Die Validität dieses Programms wurde an einer Stichprobe von 25 Psychiatern überprüft, die der Meinung waren, dass sie über einen Fernschreiber mit einem "echten" Patienten in Verbindung ständen. Aufgrund dieser Kontakte stellten

23 Psychiater die Diagnose "Paranoia"; die beiden anderen diagnostizierten einen Hirnschaden.

Ein grosser Teil von Simulationsmodellen in der Klinischen Psychologie untersucht verschiedene Dimensionen des Therapie-Gesprächs. MOSER, SEIDENSTÜCKER & SEIDENSTÜCKER (1974) nennen fünf Komponenten, die für diese Modelle kennzeichnend sind:

- Ein Wörterbuch, das das verbale Gedächtnis des Patienten und des Therapeuten darstellt.

- Zerlegungsmechanismen, die die eingehenden Sätze über das Wörterbuch syntaktisch auf ihren Gehalt hin analysieren.

- Einen als semantisches Gedächtnis konzipierten Speicher, in dem ein Modell des Gesprächspartners vorhanden ist, bzw. während des Gesprächs aufgebaut werden kann.

- Ein Verarbeitungssystem, das die eingegebenen Sätze auf ihren syntaktischen Gehalt hin untersucht.

- Eine Sammlung von Transformationsregeln, die als Reaktion auf die eingegebenen Sätze unter Berücksichtigung des Affektzustandes und des bis zu diesem Zeitpunkt über den Gesprächspartner aufgebauten Modells die Erzeugung von Antworten ermöglicht.

Mit dem nach diesen Prinzipien konstruierten Programm ELIZA (WEIZENBAUM, 1966) gelingt es z.B., den Ablauf einer nichtdirektiven Therapie zu simulieren.

Die eingehenden Klienten-Äusserungen werden zerlegt und auf bedeutsame Schlüsselbegriffe untersucht; danach wird eine entsprechende Therapeutenäusserung erstellt. Lautet die Patientenäusserung z.B. "Ich brauche Hilfe.", so wird nach den Konstruktionsregeln dieses Programms zunächst "Ich" durch "Sie" ersetzt ("Sie brauche Hilfe."). Danach analysiert das Programm das zweite Schlüsselwort "brauche", für das in Verbindung mit dem ersten Schlüsselwort "Sie" eine Reihe von Synonymen gespeichert sind. So ergibt sich die

Therapeutenäusserung: "Was würde es für Sie bedeuten, wenn Sie Hilfe bekämen." Zusätzlich vermerkt das Programm, dass diese Redewendung benutzt worden ist, so dass sie in Transformation des Begriffs "brauche" erst nach Verwendung der übrigen Synonyme erneut benutzt wird.

Wird kein Schlüsselwort gefunden, weicht das Programm auf eine unspezifische Antwort aus, die das Gespräch weiterleiten soll, oder es sucht in seinem semantischen Gedächtnis nach einem für den Patienten besonders aktuellen Thema.

Solche Programme genügen den praktischen Anforderungen noch lange nicht so sehr, dass der Therapeut durch den Computer ersetzt werden könnte. Ermutigend mag in diesem Zusammenhang aber die Beobachtung von HILF et al. (1971) sein, dass ein mutistischer Patient nach mehreren Therapiekontakten, während derer er nur über einen Fernschreiber mit dem Therapeuten in Verbindung stand, wieder zu sprechen begann.

2. DIE ROLLE DES COMPUTERS IN DER DIAGNOSTIK

Geht man von der Abfolge eines diagnostischen Prozesses aus, so lassen sich bei der praktischen Tätigkeit eines Diagnostikers grob vier Schritte unterscheiden:

- o Testdurchführung;
- o numerische Testauswertung;
- o qualitative Testinterpretation;
- o Beratung.

In diesen vier Bereichen gibt es unterschiedlich starke Ansätze zur Automatisierung: Ein Überblick über Forschungsergebnisse zur computer-unterstützten Diagnostik ist aus zwei Gründen erschwert. Zum einen macht sich das Fehlen eines übergreifenden Ansatzes für eine automatisierte Diagnostik darin bemerkbar, dass zwar zu den einzelnen Schritten eine Reihe von Untersuchungen vorliegen, diese jedoch recht beziehungslos nebeneinander stehen und nicht einfach zu einem Gesamtsystem verbunden werden können. Andererseits gibt es eine Vielzahl von Arbeiten, die in etwa unter den Begriff Testinterpretation fallen, jedoch von ihrer Anlage her nicht auf den praktischen Einsatz bei der computer-unterstützten Diagnostik zielen. Es handelt sich dabei um die lange Reihe von Untersuchungen zum diagnostischen Urteilsprozess, die direkt oder indirekt durch MEEHLs (1954) Gegenüberstellung zwischen statistischem und klinischem Vorgehen ausgelöst worden sind.

Wir wollen nun zunächst überprüfen, welcher Grad der Automatisierung in den einzelnen Bereichen des diagnostischen Prozesses bisher erreicht worden ist, und dann auf die Untersuchungen zum diagnostischen Urteilsprozess eingehen, soweit sie für eine Automatisierung der Diagnostik verwertbar erscheinen.

A. Die Automatisierung einzelner Phasen des diagnostischen Prozesses

a. Testdurchführung

Die Bedeutung einer automatisierten Testdurchführung kann zunächst in der Entlastung des praktisch tätigen Psychologen von der täglichen Routinearbeit des Testens gesehen werden, der damit erheblich mehr Zeit für Beratung und Therapie erhält. Aus praktischer Sicht ist es daher naheliegend, zunächst ein bewährtes Testverfahren soweit für eine automatisierte Durchführung aufzuarbeiten, dass der Psychologe in der Phase der Testdurchführung überflüssig wird. Diesen Weg ist man z.B. mit der Automatisierung des WECHSLER-Tests (WAIS) am Consulting Center in Columbus, Indiana, gegangen (ELWOOD, 1969, 1972a, 1972b, 1972c; ELWOOD & GRIFFIN, 1972; ELWOOD, 1973).

Zur Testdurchführung wird ein speziell konstruierter, gut beleuchteter und klimatisierter Raum benutzt. Der Proband sitzt an einem Arbeitstisch, auf dem sich Projektionsschirm, Lautsprecher, Mikrofon, mehrere Schubkästen und eine Antworttastatur befinden. Die Anweisungen für die einzelnen Untertests und die Aufgaben des Verbalteils werden über ein Tonband vorgegeben, nachdem der Proband jeweils die Taste READY betätigt hat. Die Antworten werden auf Tonband aufgezeichnet.

Das Aufgabenmaterial des Handlungsteils befindet sich in den Schubladen, die nach Betätigung der READY-Taste aufspringen. Über Kontakte wird die Zeit zwischen Herausnehmen und Zurückstellen der Schubladen registriert. Nach der Testdurchführung müssen die Antworten des Verbalteils vom Tonband abgeschrieben und nach der Originalanweisung gepunktet werden; für die Bewertung des Handlungsteils ist der Schubladeninhalt auf Richtigkeit zu überprüfen.

Das beschriebene automatisierte System ist der ursprünglichen Testform möglichst genau nachgebildet und wirkt durch die daher notwendigerweise auf mechanischem Wege vorgenommene Automatisierung zum Teil recht unbeholfen.

Auch erfordert es bei der Auswertung (Abschreiben der Tonbänder, Antwortbewertung) den Einsatz von Hilfskräften.
Jedoch konnten ELWOOD und GRIFFIN zeigen, dass selbst ein von der Durchführungsart her so komplizierter Test wie der WAIS in eine adäquate automatisierte Form übertragen werden kann:

In verschiedenen Reliabilitätsüberprüfungen mit automatisierten Formen ergaben sich für Untersuchungszwischenräume von bis zu 224 Tagen Retestkoeffizienten von $< .9$ für Verbal-, Handlungs- und Gesamt-IQ. Auch korrelierten herkömmliche und automatisierte Durchführung $< .9$ bei einem durchschnittlichen Abstand von 9 Tagen. Daneben konnte nachgewiesen werden, dass auch Versuchspersonen mit niedriger Intelligenz in der Lage sind, die automatisierte Form zu bewältigen (Korrelation zwischen beiden Formen $< .9$).

Zusätzlich kam ELWOOD (1972) in einer Kosten-Nutzen-Analyse zu dem Ergebnis, dass sich mit dem beschriebenen System die Kosten für eine Testdurchführung im Vergleich zur herkömmlichen Durchführung durch einen Psychologen um etwa die Hälfte senken lassen, wenn man dabei die Anschaffungskosten für die Maschinen nicht in Rechnung stellt.

Mit der Möglichkeit zur Übertragung eines herkömmlichen Verfahrens in eine automatisierte Form wird zwar der Praktiker entlastet, doch ergeben sich damit eine Reihe von Fragen nach den psychologischen Auswirkungen dieser Testdurchführung, die bisher nur zu einem geringen Teil abgeklärt sind.

Für die unterschiedliche Effektivität herkömmlicher gegenüber automatisierter Testdurchführung ergibt sich ein Hinweis aus dem Bereich des programmierten Lernens.

CAHOON, PETERSON & WATSON (1968) untersuchten die Effektivität zweier programmierter Unterweisungsformen in Abhängigkeit vom Interessensschwerpunkt des Lernenden. Nach dieser Arbeit erzielen technisch interessierte Versuchspersonen bessere Lernfortschritte bei der Unterweisung durch eine Lehrmaschine im Vergleich zur programmierten Unterweisung in Buchform, während sich bei literarisch Interessierten gegenteilige Befunde ergaben.

Die Leistungssteigerung sonst bei der Testdurchführung benachteiligter Probandengruppen in einem automatisierten Testverfahren konnten JOHNSON & MIHAL (1973) nachweisen.

Die Autoren gingen von dem Gedanken aus, dass in der üblichen Testsituation farbige Probanden gegenüber weissen Probanden benachteiligt sind, da sie z.B. vor dem meist weissen Untersuchungsleiter Angst haben, mit der Testsituation nicht so vertraut sind u.a.m.. Betrachtet man beide Versuchspersonengruppen gemeinsam, so ergeben sich keine Unterschiede zwischen einem in herkömmlicher Form dargebotenen Leistungstest und einer parallelen Computerform. Beim Vergleich zwischen den Versuchspersonengruppen schneiden in der üblichen Testform die weissen Probanden besser ab, während in der Computerform keine Leistungsunterschiede zwischen den beiden Versuchspersonengruppen festzustellen sind.

Von der Vermutung ausgehend, dass die automatisierte Testsituation neutraler und vertrauenswürdiger sei und deshalb objektivere Untersuchungsbedingungen ermögliche (SMITH, 1963), überprüften HEDL, O'NEIL & HANSEN das Ausmass an Testangst in beiden Untersuchungsformen.

Parallel zum SLOSSEN-Intelligence-Test (SLOSSEN, 1963) - Korrelation zwischen beiden Durchführungsformen r =.75 - wurde in der herkömmlichen und automatisierten Testsituation mit dem "State-Trait Anxiety Inventory" von SPIELBERGER, GORSUCH & LUSHENE (1970) die Testangst gemessen. Entgegen der Vermutung ergab sich, dass diese in der herkömmlichen Durchführungsform wesentlich geringer war.

In einer ähnlichen Untersuchung stellten LUSHENE, O'NEIL & DUNN (1972) fest, dass bei der automatisierten Durchführung des MMPI lediglich zu Anfang grössere Angstzustände auftreten, während gegen Ende der Untersuchung zwischen beiden Darbietungsarten keine Unterschiede mehr bestehen.

Über die Auswirkungen der automatisierten Versuchssituation auf die Beantwortung von Items in Persönlichkeitsfragebögen gibt es bisher zwei sich widersprechende Versuchsergebnisse.

Eine Untersuchung von EVAN & MILLER (1969) ergab, dass Fragen, die von den Probanden als "persönlich" empfunden wurden, in der automatisierten Durchführung offener beantwortet wurden, während bei neutraleren Fragen keine Unterschiede zwischen den Darbietungsformen auftraten.

In einer analogen Studie von KOSON et al. (1970) wurden jedoch zwischen den Durchführungsarten keine Unterschiede festgestellt.

b. Numerische Testauswertung

Verschiedene Testverlage und Datenverarbeitungsfirmen bieten die Möglichkeit einer automatisierten numerischen Testauswertung an - z.B. kann auch der IST (AMTHAUER, 1970) so ausgewertet werden. Es ist hier nicht unser Ziel, eine Bestandsaufnahme der automatisch auswertbaren Verfahren vorzunehmen - s. dazu vor allem BÜRLI (1973) - sondern wir wollen nur auf einige Grundprinzipien hinweisen.

Grundsätzlich lässt sich zumindest jeder Test in Mehrfach-Wahl - Form automatisch auswerten. Entweder müssen dazu die Probandenantworten nach der Testdurchführung auf einen maschinen-lesbaren Datenträger übertragen werden (z.B. auf Lochkarten), oder sofort vom Probanden auf einem entsprechenden Antwortblatt, das z.B. von einem optischen Markierleser verarbeitet werden kann, angestrichen werden.

Die Dienste einer Auswertungsstelle bestehen dann vor allem in der Berechnung der Roh- und Standardwerte, in der Gegenüberstellung der Individual- oder Gruppentestwerte mit den Werten einer bestimmten Vergleichspopulation. Weiter gibt es die Möglichkeit, bestimmte Verteilungen berechnen und nach Aussenkriterien geordnete Listen erstellen zu lassen. Auch können die verschiedenen Paramter der Itemanalyse berechnet werden.

Die Auswertungsobjektivität eines Tests wird durch die automatisierte Auswertung optimal erhöht; psychologisch können jedoch verschiedenartige Antwortblätter unterschiedlich auf die Probanden wirken (vgl. BELL, HOFF & HOYT, 1964), was bei der qualitativen Testinterpretation

unbedingt zu berücksichtigen ist.

Darüberhinaus gibt es Bemühungen, Schulaufsätze und Kurzabhandlungen (BHUSHAN & GINTHER, 1968; PAGE, 1968), thematische Testverfahren (SMITH, 1968) und Satzergänzungstests (GOLDBERG, J.B., 1966, VELDAMN 1967) automatisch auszuwerten, wobei man meist von dem GENERAL-INQUIRER-Programm (z.B. STONE et al., 1967) ausgeht.

c. Qualitative Testinterpretation

Mit LEVY (1963) lässt sich die Testinterpretation als Rekonstruktion und Erklärung einer Situation durch eine alternative Beschreibung auffassen. Dabei werden die aus einem Test im weitesten Sinne stammenden Daten in die Fachsprache der Psychologie übersetzt und in einen psychologisch möglichst plausiblen Zusammenhang gestellt. Dies geschieht - oder sollte zumindest geschehen - auf dem Hintergrund einer expliziten psychologischen Theorie. Je weiter die psychologische Theorie eines Testverfahrens expliziert ist, desto eher kann die Aufgabe der Testinterpretation einem Automaten übertragen werden.

Die meisten Systeme für eine automatisierte Testinterpretation sind für den MMPI ('Minnesota Multiphasic Personality Inventory', HATHAWAY & McKINLEY, 1967) entwickelt worden. Dies lässt sich auf der einen Seite aus der breiten und häufigen Anwendung dieses Verfahrens begründen, zum anderen spielt es aber auch sicher eine gewichtige Rolle, dass MEEHL, der die Diskussion über klinische und statistische Urteilsbildung ausgelöst hat, von der Minnesota - Universität stammt.

Wir wollen nun einen Überblick über die wichtigsten Systeme zur automatischen Interpretation des MMPI geben (vgl. BÜRLI, 1972) und danach beispielhaft auf ein System

zur Interpretation verschiedenartiger psychologischer Daten eingehen.

Eines der ersten Interpretationsprogramme wurde an der Mayo-Klinik entwickelt. Für jede Subskala gibt es fünf verschiedene Interpretationsstufen, denen je nach Skalenwert eine bestimmte psychologische Feststellung (Statement) zugeordnet ist. Zur Vermeidung widersprüchlicher Beschreibungen und für die gemeinsame Verarbeitung auffälliger Werte in mehreren Skalen enthält das Programm zusätzlich elf Verzweigungsoperationen. Bei der sprachlichen Formulierung der einzelnen Statements wurde besonders darauf geachtet, dass mit Rücksicht auf die verschiedenen Empfänger der automatisch erstellten Gutachten diagnostische Abqualifizierungen wie "Schizophrenie" oder"Psychopathie"vermieden und die Gutachten insgesamt von Patienten und Klinikern anerkannt werden konnten. Zunächst wurde dieses System bei etwa 40% der Patienten in der Mayo-Klinik angewandt, so dass bis 1965 schon rund 60.000 Testprotokolle vorlagen (SWENSON et al., 1965); heute zählt es zu den Routineverfahren der diagnostischen Abteilung.

Eine Übertragung des Mayo-Systems auf die besonderen Anforderungen einer psychiatrischen Klinik wurde von GLUECK & REZNIKOFF (1965) vorgenommen, indem sie bei der automatischen Interpretation stärker den Bezug zur Psychopathologie herstellen. Dieses System enthält 70 Interpretationsstatements, von denen sich 60 auf 14 MMPI-Skalen beziehen, während die übrigen Statements der Interpretation bestimmter Konfigurationen dienen (FOWLER, 1969).

Das System von FOWLER wurde mit dem Ziel konzipiert, klinischen Psychologen und Psychiatern einen klinisch möglichst brauchbaren Bericht an die Hand zu geben, der sich von einem "per Hand" erstellten Gutachten möglichst wenig unterscheiden sollte. Während einer vierjährigen Entwicklungsphase wurden von FOWLER (1964, 1967, 1969; FOWLER et al., 1968a) ständig Verbesserungen aufgrund von etwa 2.000 Protokollen vorgenommen. Von dem Programm werden zehn klinische Skalen, vier Validitäts- und eine Reihe von Spezialskalen berücksichtigt, sowie für einzelne Skalen und Skalenkonfigurationen Interpretationen vorgenommen. Auch nach der Übernahme dieses Systems für die kommerzielle MMPI-Interpretation wurde es ständig weiter verbessert und überprüft. (FOWLER & MILLER, 1969; WEBB, 1970; WEBB, MILLER & FOWLER, 1970).

FOWLER selbst (FOWLER, 1967a, 1968; FOWLER & COYLE, 1968; FOWLER et al. 1968b) berücksichtigt meist Studentenstichproben, da er sein Programm vor allem für die Studentenberatung einsetzt. - Bei der im deutschsprachigen Bereich angebotenen kommerziellen MMPI-Interpretation[1] handelt es sich um eine modifizierte Fassung des hier beschriebenen Systems. (GEHRING & BLASER, 1972)

Weiter sind noch die Systeme von FINNEY (1967) und DUNLOP (1966) zu nennen, doch ist eine vollständige Aufzählung aller automatisierten MMPI-Systeme kaum möglich. - Natürlich gibt es auch Programme zur automatischen Interpretation anderer Persönlichkeitsfragebögen, wie etwa für das 'California Psychological Inventory' (FINNEY, 1966) und eine Reihe von Tests von CATTELL, die vom IPAT-Institut angeboten werden.

Besonders im psychiatrischen Bereich werden Versuche unternommen, nicht nur die Ergebnisse eines einzelnen Testverfahrens, sondern eine Fülle unterschiedlicher Daten automatisch auszuwerten. Als Beispiel können wir hier das 'Missouri Mental Status Examination' System anführen. (SLETTEN et al. 1970a, 1970b, 1971, 1972; EATON et al. 1970)

Aufgrund von über 200 verschiedenen Daten - Angaben zur Anamnese, Befragung der Angehörigen, Berichte der Ärzte und der Pfleger, Informationen aus Laboruntersuchungen, Fragen aus Persönlichkeitsinventaren - werden vom System diagnostische und prognostische Berichte erstellt: wahrscheinliche Verweildauer des Patienten, Suizidgefährdung, Ansprechen auf verschiedene Behandlungsarten oder die Wahrscheinlichkeit der sozialen Wiedereingliederung.

---

1) Die maschinelle MMPI-Auswertung wird von der Firma HOFFMANN LA ROCHE (Basel) in Zusammenarbeit mit dem HUBER - Verlag (Bern) angeboten. - Voraussetzung für diese Form der Auswertung ist jedoch, dass die Probandenantworten auf einem speziellen (maschinen-lesbaren) Antwortblatt markiert sind.

Den Ausgangspunkt für Diagnose und Prognose bildet der Vergleich mit den Krankengeschichten von rund 5.000 Patienten. In statistischen Vergleichen ergaben sich als die fünf wichtigsten Prediktoren: Alter und Geschlecht des Patienten, Hinweise auf organische Hirnerkrankungen, paranoide Wahnvorstellungen und der Grad der sozialen Isolation. Bei Benutzung dieser Kriterien ergab sich die Übereinstimmung mit psychiatrischen Diagnosen in 43% der Fälle; wurden insgesamt 32 Variablen berücksichtigt stieg die Übereinstimmung auf 54%. SLETTEN et al. (1972) führen aus, dass diese Übereinstimmung etwa der zwischen zwei Psychiatern entspricht.

In diesem Zusammenhang ist auch das Programm DIAGNO (SPITZER & ENDICOTT, 1968, 1969) zu nennen, das von der Messung psychologischer Standard-Variabler (einschliesslich des MMPI) und strukturierten psychologischen Interviews ausgeht. PLATMAN & WEINSTEIN (1970) berichten von einer fast vollständigen Übereinstimmung zwischen Psychiater- und Computerdiagnosen, wenn zur Klassifikation der Probanden nur die Kategorien "gesund", "neurotisch" und "psychotisch" benutzt werden.

d. Beratung

Der Bereich computer-unterstützter Beratung wird z. Zt. nur auf dem engeren Gebiet der Studien- und Berufsberatung praktisch angegangen. HARRIS (1974) kommt auf etwa 30 verschiedene Beratungssysteme, die in den letzten zehn Jahren in den USA entwickelt worden sind. Von diesen gehören jedoch zwanzig zu den Systemen mit indirekter Befragung; d.h. der Proband kreuzt auf einem Fragebogen die ihn interessierenden Studien- oder Berufsbereiche an und erhält später von der - meist kommerziellen - Auswertungsstelle eine Vorschlagsliste, die seinen Wünschen möglichst nahe kommen soll.

Bei den direkten Systemen - der Proband steht über ein Terminal in ständigem Kontakt mit dem Computer - kann man zwischen Programmen, die nur den Rückgriff auf die Vielfalt der gespeicherten Daten gestatten, und solchen

unterscheiden, die stärker an das Vorgehen eines menschlichen Beraters angelehnt sind. In der letzteren Gruppe werden auf der einen Seite möglichst vielfältige Informationen über den Probanden (Testdaten usw.) herangezogen; vor allem aber ist es für diese Gruppe kennzeichnend, dass der Beratungsprozess im Sinne einer sequentiellen Entscheidung vor sich geht, und der Benutzer ständig Rückmeldungen über den bisher zurückgelegten Entscheidungsweg erhält. Bei der Überprüfung der Effektivität solcher Systeme konnte auch in follow-up-Studien festgestellt werden (PRICE & GILMORE, 1973; PRICE, 1974), dass keine Unterschiede zwischen herkömmlicher und automatisierter Beratung bestehen.

Eine Sonderstellung nimmt in diesem Zusammenhang das 'Harvard Information System for Vocational Decisions' (ELLIS & TIEDEMANN, 1970; SUPER, 1970) ein, bei dem zum Teil versucht wird, die Prinzipien einer nichtdirektiven Beratung zu verwirklichen. Doch muss dieses System auch heute noch als ein Modell für die Zukunft angesehen werden.

B. DER DIAGNOSTISCHE URTEILSPROZESS

Es ist unmöglich, an dieser Stelle auch nur einen gewissen Überblick über die Fülle der Untersuchungen zum diagnostischen Urteilsprozess zu geben - eine Lawine, die direkt oder indirekt durch MEEHL (1954) ausgelöst worden ist. Auf dem Hintergrund einer automatisierten Diagnostik geht es hier nicht darum, die wissenschaftstheoretischen Standpunkte der verschiedenen Schulen gegeneinander abzuwägen, sondern wir sehen die einzelnen Untersuchungen als Grundlagenforschung für den Bereich einer automatisierten Testinterpretation an.

Mit GREEN (1968) unterscheiden wir zwei Hauptrichtungen

in der Forschung zum diagnostischen Urteilsprozess: "Blackbox"-Methoden und die Methode der Protokoll-analyse, die in etwa MEEHLs Kategorien der statistischen vs. klinischen Vorgehensweise entsprechen.

Die "Black-box"-Strategie war vor allem für behavioristisch ausgerichtete Psychologen lange Zeit die Forschungsmethode der Wahl; es wird nicht untersucht, warum aufgrund bestimmter Daten eine bestimmte Entscheidung getroffen wird, sondern es interessiert allein der - mathematisch - beschreibbare Zusammenhang zwischen eingegebenen und ausgegebenen Daten.

Die Methode der Protokoll-Analyse, die versucht, die "Black-box" zu öffnen und den in ihr arbeitenden Mechanismus aufzudecken, wurde lange als "Introspektion" angesehen und daher als "nicht wissenschaftlich" abgelehnt.

Wir wollen für jede dieser beiden Richtungen einen typischen Vertreter vorstellen und dann auf ein Modell eingehen, das nicht zu beschreiben versucht, wie Diagnostiker vorgehen, sondern ihnen vorschreibt, wie sie vorzugehen haben.

a. 'Black-box'-Methoden

Ein grosser Teil der Untersuchungen unter diesem Forschungsansatz stammt aus dem 'Oregon Research Institute' oder wurde von dort aus angeregt (GOLDBERG, 1970; HOFFMANN, 1968; WIGGINS, 1971). Hier wird die Überlegenheit der aktuarischen gegenüber der klinischen Vorgehensweise als erwiesen angesehen (GOLDBERG, 1965; WIGGINS & KOHEN, 1971).

Um zu beschreiben, wie es zu einer bestimmten klinischen Entscheidung kommt, werden die Entscheidungen des Diagnostikers als abhängige Variablen und die Daten, aufgrund

derer die Entscheidungen getroffen werden, als unabhängige Variablen angesehen (HOFFMANN, 1960; NAYLOR & WHERRY, 1965). Dabei benutzen die Autoren als "unabhängige Variablen" ausschliesslich Skalenwerte des MMPI. Die multiple Regression zwischen Testdaten und Urteilen kann dann als "lineares Modell" des Klinikers angesehen werden (GOLDBERG, 1968, 1970). Die Korrelation zwischen Kliniker-Urteil und Kriterium beschreibt die Genauigkeit des Urteils; zwar werden hier als Aussenkriterium wiederum "Kliniker-Urteile" benutzt, doch stand diesen"Klinikern" als den behandelnden Ärzten eine Fülle weiterer Informationen zur Verfügung. In den angeführten Untersuchungen konnte gezeigt werden, dass die aufgrund solcher Modelle erstellten Programme sowohl dem einzelnen Kliniker als auch dem durchschnittlichen Urteil mehrerer Kliniker überlegen sind.

Der Nachweis einer guten linearen Repräsentation des klinischen Urteils durch eine multiple Regressionsgleichung wurde von verschiedenen Untersuchern als zwar die Daten repräsentierend aber psychologisch nicht befriedigend angesehen (EINHORN, 1970, 1971). Einerseits geht diese Kritik von der Annahme aus, gerade das nicht lineare Vorgehen - die Beachtung bestimmter Konfigurationen - sei die besondere Stärke des Klinikers. Auf der anderen Seite muss man aber auch annehmen, dass ein lineares Modell, das zwar mathematisch einfach ist, nun nicht auch unbedingt kognitiv einfach sein muss, und umgekehrt das mathematisch komplexere (nicht-lineare) Modell kognitiv nicht so schwierig erscheinen mag.

Ausgehend von der Feststellung, dass vor allem bei der Interpretation projektiver Verfahren aber auch in Zusammenhang mit Profilinterpretationen vom Diagnostiker nicht-lineare Vorgehensweisen gefordert werden, untersuchte SCHULZ (1975) die mathematische Komplexität von Urteilsmodellen und ihrer kognitiven Wirkung. Es zeigte sich, dass nicht allzu hohe Erwartungen an nicht-lineare Vorgehensmöglichkeiten gestellt werden können. Denn zum einen lässt sich ein grosser Teil nicht-linearer Beziehungen gut durch lineare Regressionsgleichungen approximieren; damit fällt der Gewinn bei diesem Vorgehen kaum ins Gewicht. Vor allem ist aber bei nicht-linearem Vorgehen die konsequente Einhaltung der Lösungsstrategie kaum möglich, so dass der geringe potentielle Gewinn auf Grund der erhöhten Fehleranfälligkeit kaum realisiert werden kann.

b. Protokoll-Analysen

Die Methode der Protokoll-Analyse zur Erforschung des diagnostischen Urteilsprozesses wird vor allem von B. KLEINMUNTZ benutzt. Dabei handelt es sich fast immer um die aufgrund von MMPI-Daten erstellte Differentialdiagnose zwischen "angepassten" und "unangepassten" Studenten.

Zur Gewinnung eines ersten Simulationsprogrammes ging KLEINMUNTZ (1963, 1963a, 1963b, 1963c) so vor, dass die MMPI-Profile von 126 Studenten, die nach einem Aussenkriterium als angepasst bzw. unangepasst eingestuft worden waren, von zehn MMPI-Experten unter dieser Fragestellung sortiert werden sollten. Der -gemessen am Aussenkriterium- erfolgreichste Diagnostiker hatte dann seinen Sortiervorgang zu wiederholen und dabei laut zu kommentieren. Das aufgrund dieses Protokolls erstellte Programm war nach verschiedenen Korrekturen schliesslich dem Experten überlegen.

Dieses Vorgehen wurde von KLEINMUNTZ (1967, 1968, 1969, 1969a, 1970) mehrfach überprüft. Dabei ergab sich wiederum eine Überlegenheit des Simulationsprogramms sowohl gegenüber dem Urteil des besten Klinikers als auch gegenüber den durchschnittlichen Kliniker-Urteilen. - Bei einem Vergleich mit dem MMPI-Programm von FOWLER (1968b) erbrachten beide Programme bezogen auf ein Aussenkriterium ähnlich valide Diagnosen.

c. Ein normatives Modell psychologischen Urteilens

Mit seiner Zielsetzung, eine Lehre vom korrekten Diagnostizieren zu entwickeln, steht WESTMEYER (1972) mit seinem normativen Ansatz im Gegensatz zu den bisher beschriebenen Ansätzen des diagnostischen Urteilsprozesses, die er als deskriptiv, explanativ oder spekulativ einordnet. WESTMEYER entwickelt ein weitgehend formalisiertes Modell, wobei sich sein Entwurf auf logische und wissenschaftstheoretische Analysen einzelner Aspekte dieses Prozesses stützt.

Grundlegend für seinen Entwurf ist der logisch-systematische Begriff der wissenschaftlichen Erklärung, der in

Übereinstimmung mit dem psychologischen Begriff Diagnose gesehen wird. Als Schema der wissenschaftlichen Erklärung dient das für den deterministischen Fall geltende HEMPEL-OPPENHEIM-Schema der deduktiv-nomologischen Erklärung. Denn bei der Diskussion anderer Inferenztheorien, insbesondere des Modells von SARBIN, TAFT & BAILEY (1960) wird nachgewiesen, dass der dort benutzte statistische Syllogismus zu logischen Widersprüchen führen kann und deswegen als Modell ungeeignet ist. Ausgangspunkt für eine Diagnose zu einem Zeitpunkt ist das Rationale Corpus RC_t zur Zeit t, die Gesamtheit der zu diesem Zeitpunkt vorhandenen psychologischen Gesetze. Der eigentliche diagnostische Prozess vollzieht sich dann in zwei Bereichen: Für ein zu erklärendes Ereignis sind die relevanten Gesetzmässigkeiten aus RC_t auszuwählen und die möglichen Erklärungen bereitzustellen; dann sind diese potentiellen Diagnosen - operationalisierbar als Testergebnisse im weitesten Sinne - auf ihr tatsächliches Zutreffen hin zu überprüfen. Für die Aufgaben der Auswahl und Prüfung werden Algorithmen entwickelt, wobei auch die Abfolge der einzelnen Prüfschritte durch einen Prozess-Algorithmus bestimmt ist.

Obwohl es sich bei dieser Arbeit um einen theoretischen Entwurf handelt, haben wir ihn hier vorgestellt, weil damit die Frage, ob eine automatisierte Diagnostik überhaupt möglich ist, prinzipiell bejaht werden kann. "Ob eine automatisierte Diagnostik auch realisierbar ist, kann nicht in Diskussionen entschieden werden. Man muss vielmehr in praktischen Versuchen ermitteln, wie weit der Ansatz trägt. Übertriebener Skeptizismus behindert nur die Forschung und verzögert eine begründete Stellungnahme zu der Streitfrage." (WESTMEYER, S. 140)

3. ZUSAMMENFASSUNG UND FOLGERUNGEN

Anhand einer Reihe typischer Beispiele haben wir versucht, einen gewissen Überblick über die Einsatz- und Leistungsmöglichkeiten des Computers in verschiedenen Bereichen der Psychologie zu geben. Wenn man das grosse Gebiet der mathematisch/statistischen Bearbeitung von Daten ausklammert, lässt sich der Computer zunächst als genauer und standardisierter Experimentator auffassen.

Neben der Textausgabe auf verschiedene Medien eröffnet vor allem die Bildschirmausgabe graphischer Darstellungen, einschliesslich der Ausgabe bewegter Bilder, eine Vielzahl von Untersuchungsmöglichkeiten. Auch reichen die Möglichkeiten bei der Registrierung von Versuchspersonenreaktionen bis hin zur genauen Zeitmessung. Weiter können die im Experiment anfallenden Daten mit einer sehr geringen Fehlerquote auf eine Reihe von Datenträgern ausgegeben bzw. gleich nach Versuchsende verrechnet werden.

Über die reine Reizdarbietung und Reaktionsregistrierung hinaus lassen sich die Möglichkeiten des Computers noch effektiver nutzen, wenn parallel zur eigentlichen Versuchsdurchführung die anfallenden Daten analysiert werden, und aufgrund dieser Analyse der Versuchsfortgang gesteuert wird. Beispiele finden sich im Bereich der Wahrnehmungspsychologie und auf dem Gebiet der computer-unterstützten Unterweisung.

Im Bereich der Computersimulation haben zur Zeit solche Simulationsprogramme die grösste Bedeutung, die primär zur Lösung praktischer Probleme entwickelt wurden, aber gleichzeitig als Modelle bestimmter psychischer Funktionen dienen können. Hier sind Programme zum Erkennen von Klartext, dem Verarbeiten von Texten in natürlicher Sprache, vor allem aber aus den verschiedensten Gebieten des Problemlösens zu nennen.

Für die hier näher dargestellten Teilbereiche diagnostischer Tätigkeit (Testdurchführung, Testauswertung, Testinterpretation und Beratung) gibt es auf der praktischen Ebene mehr oder weniger weit fortgeschrittene Automatisierungsversuche, doch fehlen zum Teil die theoretischen Grundlegungen.

Bei der Testdurchführung beschränkt man sich bisher auf die Übernahme bereits bestehender Verfahren und die Untersuchung der affektiven Auswirkungen dieser Durchführungsart.

Die numerische Testauswertung ist bisher am stärksten automatisiert und bedarf aufgrund der nur nach herkömmlichen Prinzipien konstruierten Tests keiner eigenen Theorie.

Für die qualitative Testinterpretation bestehen theoretische Ansätze im Bereich der Forschung zum diagnostischen Urteilsprozess. Doch bleiben diese fast ausschliesslich auf beschreibender oder erklärender Ebene, so dass sie keinen wesentlichen Fortschritt ermöglichen. Von der Theorie her stärker befriedigende Ansätze erfordern noch eine solche Fülle von Vorarbeiten, dass sie zur Zeit kaum in der Praxis verwirklicht werden können.

Die Systeme zur automatisierten Beratung sind fast nur auf die Vermittlung von Sachinformationen angelegt, ohne dabei die Auswirkungen der Beratungsvermittlung auf den Beratenen zu berücksichtigen. Versuche, Theorien der Beratung als Grundlagen für Beratungsprogramme zu verwenden, haben bisher nur Modellcharakter.

Gehen wir nach diesem Überblick noch einmal zu der eingangs getroffenen Wertung zurück, dass im Vergleich zu den prinzipiell durch den Computereinsatz eröffneten Möglichkeiten die bis heute erarbeiteten praktischen Lösungen noch recht unbefriedigend wirken, so können wir feststellen:

Für die verschiedenartigsten Einzelprobleme finden sich Ansätze zur Automatisierung, die zum Teil so geschickt und durchdacht aufgebaut sind, dass die Arbeit des Computers in einem Sinn "intelligent" wirkt, wie man es eigentlich nur von Menschen erwartet. Doch handelt es sich dabei nur um die Lösung von eng umschriebenen Einzelproblemen, die relativ beziehungslos nebeneinander stehen und sich noch nicht zu einem Gesamtsystem verbinden lassen. Es können zumindest zwei Gründe angeführt werden, warum man bis heute noch nicht diesen Stand überwunden hat.

Auf der einen Seite ist das Fehlen einer übergreifenden Theorie des diagnostischen Prozesses zu nennen. Wie wir gezeigt haben, ist zwar die gesamte Automatisierung prinzipiell denkbar, doch gibt es noch keinen Ansatz, der in naher Zukunft in die Praxis umgesetzt werden könnte.

Die grösste Schwierigkeit besteht zur Zeit aber darin, dass die für ein System erarbeiteten Lösungen nicht einfach in ein anderes System übertragen werden können. Wenn z.B. für einen bestimmten Schritt des diagnostischen Prozesses die Analyse der natürlichen Sprache erforderlich ist, so lassen sich im Bereich der Sprachpsychologie entsprechende Arbeiten finden. Doch ist aufgrund der unterschiedlichen Zielsetzungen die Übertragung der dort gefundenen Lösungswege - wenn überhaupt - nur mit einem immensen Arbeitsaufwand möglich.

Allgemeiner bedeutet das, dass für die praktische Arbeit mit automatisierten Systemen zwar die Prinzipien bisher gefundener Lösungen übernommen werden können. Doch sind meist schon die technischen Anforderungen bei unterschiedlichen Fragestellungen so verschieden, dass praktisch der gesamte Lösungsweg neu zu erarbeiten ist.

DER DIAGNOSTISCHE GEWINN AUFGRUND DER AUTOMATISIERTEN DURCHFÜHRUNG VON TESTVERFAHREN

Bei den vier grösseren Teilbereichen des diagnostischen Prozesses (Testdurchführung, numerische Testauswertung, Testinterpretation, Beratung) ist bisher nur die Automatisierung der numerischen Testauswertung so weit entwickelt, dass ihr für die diagnostische Praxis eine grössere Bedeutung zukommt. Während für die Automatisierung von Testinterpretation und Beratung noch eine Fülle grundlegender Arbeiten zu leisten sind, gilt für die automatisierte Durchführung von Testverfahren, dass sie in absehbarer Zeit in grösserem Umfang in die Praxis einbezogen werden kann.

Aufgrund dieser praktischen Bedeutung wollen wir uns im Rahmen der vorliegenden Arbeit im Bereich der computerunterstützten Diagnostik mit dem engeren Gebiet der automatisierten Testdurchführung beschäftigen. Dabei soll hier der Nachweis erbracht werden, dass die automatisierte Durchführung eines Testverfahrens der herkömmlichen Durchführung vor allem deswegen überlegen ist, weil sie "bessere" diagnostische Informationen ermöglicht. Wir beschränken uns dabei beispielhaft auf Verfahren zur Leistungsmessung und wollen in diesem Abschnitt den Nachweis der "besseren" diagnostischen Information für solche Tests erbringen, die nach den Prinzipien der klassischen Testtheorie konstruiert sind.

Was unter "besserer" diagnostischer Information verstanden wird, ist im folgenden näher zu erläutern.

I. DAS ÜBERGREIFENDE UNTERSUCHUNGSZIEL

So lange die automatisierte Testdurchführung auch nur die in der herkömmlichen Testdarbietung übliche Richtig/Falsch-Auswertung ermöglicht, lässt sich die Notwendigkeit einer automatisierten Darbietung, die für ihre Entwicklung im Vergleich zur herkömmlichen Testdurchführung einen erheblichen Aufwand erfordert, allein aus Kosten-Nutzen-Analysen (z.B. ELWOOD, 1972a) begründen. Doch die Vorteile dieser Durchführungsart nur im ökonomischen Bereich zu sehen, hiesse, die vielfältigen, durch den Computereinsatz gebotenen, Möglichkeiten bei weitem nicht aufzugreifen.

Die computer-unterstützte Testdurchführung kann leicht zusätzliche diagnostische Informationen liefern (BÜRLI, 1975). GREEN (1970) bezweifelt, dass die ein-Bit-Information, die sich aus der Richtig/Falsch-Auswertung einer Testaufgabe ergibt, den vielfältigen, zur Lösung einer Aufgabe erforderlichen Denkprozessen des Probanden gerecht wird.So könnten z.B. als zusätzliche diagnostische Informationen die bei der Beantwortung einer Testaufgabe messbaren Latenzzeiten, die genauere Hinterfragung von Falsch-Lösungen u.a.m. herangezogen werden - vgl. ELWOOD, 1969; GREEN, 1970. Doch handelt es sich hier fast immer um Spekulationen darüber, welche Faktoren als zusätzliche Informationsquellen herangezogen werden könnten. Praktische Ergebnisse liegen bis auf wenige Ausnahmen bisher noch nicht vor.

Mit dem empirischen Teil dieser Arbeit verfolgen wir daher das Ziel, einen Teil der bei der automatisierten Testdurchführung anfallenden Informationen dahingehend zu überprüfen, ob sie im diagnostischen Sinn einen zusätzlichen Gewinn bedeuten. Beispielhaft sollen zwei Variablen auf ihre diagnostische Bedeutung untersucht werden:

- Die zur Beantwortung einer Testaufgabe benötigte Zeit;
- die von den Probanden eingeschätzte Sicherheit für die Richtigkeit ihrer Antwort.

Es ist hier anzumerken, dass mit der Untersuchung dieser beiden Variablen die Möglichkeiten zur zusätzlichen Informationsgewinnung bei einer automatisierten Testdurch_führung bei weitem nicht erschöpft sind. Vielmehr soll mit der Untersuchung dieser Variablen exemplarisch auf die bei der automatisierten Testdurchführung anfallenden besseren diagnostischen Informationen hingewiesen werden.

Unter besserer diagnostischer Information verstehen wir dabei die Erhöhung der Messgenauigkeit eines nach den Regeln der klassischen Testtheorie konzipierten Verfahrens. Wir beziehen uns also nur auf den Bereich der Reliabilität; eine Verbesserung der diagnostischen Information im Sinne einer valideren Aussage wird nur insofern erwartet, als eine hohe Reliabilität als notwendige aber nicht hinreichende Bedingung für eine gute Validität anzusehen ist - vgl. z.B. LIENERT (1969).

Bevor wir jedoch den zusätzlichen diagnostischen Informationswert einer automatisierten Testdurchführung überprüfen können, ist zunächst nachzuweisen, dass die aufgrund automatisierter Testdurchführung gewonnenen Testergebnisse denen aus einer herkömmlichen Testdurchführung entsprechen; d.h. es ist die Frage abzuklären, ob die computer-gesteuerte Testdarbietung die herkömmliche Darbietung zumindest hinreichend ersetzen kann.

Daraus ergibt sich die Frage, wie bei der Übertragung eines herkömmlichen Testverfahrens in eine automatisierte Form vorzugehen ist, damit die herkömmliche Fassung durch die automatisierte ersetzt werden kann. Ein erstes Untersuchungsziel liegt also in der Entwicklung solcher Übertragungsregeln.

Zuvor ist jedoch noch durch die Bereitstellung eines entsprechenden technischen Systems eine weitere notwendige Bedingung zur Automatisierung eines Testverfahrens zu erfüllen.

Um das Ziel dieser Untersuchung, den Nachweis einer genaueren diagnostischen Information bei automatisierter Testdurchführung, erreichen zu können, müssen wir also in drei, einander folgenden, Schritten vorgehen:

1. Ausgehend von den vorhandenen technischen Möglichkeiten ist ein Programmsystem zur automatisierten Testdarbietung zu entwickeln. (s. Teil II.)

2. Es sind verschiedene Regeln zur Übertragung eines herkömmlichen Testverfahrens in eine automatisierte Form zu entwickeln. Danach ist empirisch zu überprüfen, aufgrund welcher Übertragungsregeln die Ergebnisse aus herkömmlicher und automatisierter Darbietung sich möglichst entsprechen. (1. Versuchsserie; s. Teil III.)

3. Mit einem, nach den günstigsten Übertragungsregeln entwickelten, automatisierten Testverfahren ist dann die Überlegenheit der automatisierten gegenüber der herkömmlichen Testdurchführung nachzuweisen. (2. Versuchsserie; s. Teil IV.)

II. ENTWICKLUNG DER TECHNISCHEN UND PROGRAMMTECHNISCHEN GRUNDLAGEN EINES SYSTEMS ZUR AUTOMATISIERTEN TEST-DURCHFÜHRUNG

In gewissem Sinne ist die Feststellung trivial, dass zur Durchführung eines computer-gesteuerten Tests bestimmte "technische" Voraussetzungen geschaffen sein müssen. Doch bedeutet dies im Bereich der elektronischen Datenverarbeitung, dass nicht nur eine bestimmte Anzahl von Geräten zur Verfügung stehen muss, sondern dass über diese "Hardware" hinaus auch die entsprechenden Steuerungsprogramme, allgemeiner die "Software", benötigt werden. Vor dem eigentlichen Beginn mit Versuchen zum computer-gesteuerten Testen gilt es also zunächst, aufgrund der vorhandenen (technischen) Möglichkeiten ein System zu entwickeln, mit dem Tests automatisch dargeboten werden können.

1. "HARDWARE"

Die "serienmässige" Computerausstattung des Psychologischen Instituts der Universität Münster - vgl. LANGHEINRICH (1975) - besteht aus einer Zentraleinheit IBM 1131 mit einer Kernspeicherkapazität von 8.192 Worten zu 16 Bits. Die Speicherkapazität wird durch eine Magnetplatte mit 514.000 Worten wesentlich erweitert, wobei für die Magnetplatte natürlich eine wesentlich längere Zugriffszeit besteht. Die Ein- und Ausgabe von Daten auf Lochkarten erfolgt über einen Lochkarten-Lesestanzer IBM 1442. Weitere Eingabemöglichkeiten bestehen über die Konsol-Tastatur und, vor allem zur Programmsteuerung über 16 Konsolschalter. Die Datenausgabe erfolgt normalerweise über einen Zeilendrucker IBM 1132; es kann aber auch die wesentlich langsamere Konsolschreibmaschine oder ein über die logische Nummer des Lochstreifenstanzers angesprochenes Datensichtgerät INFOTON angesprochen werden.

Im Prinzip sind damit die technischen Voraussetzungen für einen Dialog zwischen Mensch und Computer gegeben: der Computer stellt seine Fragen über die Konsolschreibmaschine oder den Bildschirm, der Benutzer antwortet über die Konsoltastatur, die Ergebnisse werden über Drucker oder Stanzer ausgegeben.

Vor allem um äussere Störeinflüsse auszuschalten, ist es jedoch notwendig, einen computer-gesteuerten Versuch ausserhalb des eigentlichen Rechnerraums durchzuführen. Technisch bieten sich dazu Datensichtgeräte mit einer Kabelverbindung zum Rechner an.

Im vorliegenden Fall war es nun das einfachste, über normale FORTRAN-Schreibbefehle ein Datensichtgerät INFOTON anzusprechen, das aber selbst noch nicht als Datenausgabestation diente, sondern nur zur Generierung der alphanumerischen Zeichen für die Bildschirmausgabe benutzt wurde. Die eigentliche Bildschirmausgabe erfolgte dann über einen Fernseh-Monitor im Versuchsraum. Dieser Zwischenschritt war notwendig, da zum Zeitpunkt der Versuchsdurchführung zwar die notwendigen Programme für die Zeichenausgabe aber noch nicht für die Dateneingabe vorlagen.

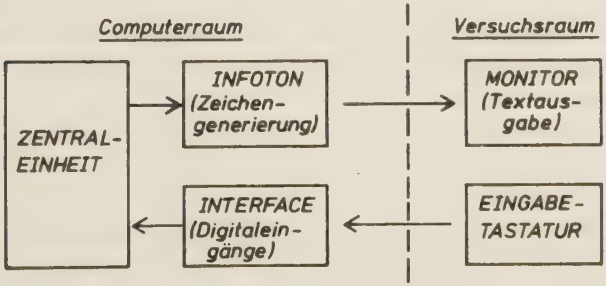


Abb. 1: Schematische Darstellung des äusseren Versuchsaufbaus zur automatisierten Testdurchführung.

Zur Dateneingabe stand der Versuchsperson eine numerische Tastatur mit drei zusätzlichen Funktionstasten ("Start", "Eingabe-Ende", "Eingabe-Korrektur") zur Verfügung. Über die Digitaleingänge des "Interface"(einer Schaltstelle zwischen dem Computer und peripher angeschlossenen Zusatzgeräten) wurden die eingetasteten Werte dann in für den Computer geeignete Impulse umgewandelt und zu ihm übertragen (vgl. Abb. 1).

Zusätzlich zu den beschriebenen Einheiten wurde zur Zeitmessung noch das Zeitregister des "Interface" benötigt. Dabei handelt es sich nicht um eine Uhr im eigentlichen Sinn, sondern um einen Impulszähler, dessen Zeitintervalle vom Computer aus festgelegt werden können, und dessen Inhalt bei Bedarf abgerufen oder gelöscht werden kann.

Nach dieser Beschribung des äusseren Aufbaus für einen Versuch im Dialog mit einem Computer, ist hier kurz aufzuzeigen, dass abgesehen von den programmtechnischen Schwierigkeiten, auf die im nächsten Abschnitt einzugehen ist, ein computergesteuerter Versuch eine Fülle technischer Schwierigkeiten mit sich bringt - vgl. LANGHEINRICH (1975).

So liegt zwischen Versuchsplanung und Versuchsdurchführung vor allem bei komplexeren Versuchen aufgrund des erforderlichen Programmieraufwandes eine recht lange Zeit, in der z.B. nach der Anschaffung leistungsfähigerer Geräte eine ständige Veränderung des äusseren Versuchsaufbaus und damit verbunden auch eine programmtechnische Umstellung erforderlich wird.

Für die ersten Versuche zur automatisierten Testdurchführung stand z.B. kein Datensichtgerät zur Verfügung, das im oben beschriebenen Versuchsaufbau die Generierung der alphanumerischen Zeichen übernimmt, sondern die Textausgabe erfolgte über einen Oszillographen, dessen 512*512 Rasterpunkte vom Computer aus einzeln angesprochen wurden. Die Textausgabe bestand dann aus folgenden Schritten: Für ein auszugebendes Zeichen waren die gespeicherten Koordinaten der einzelnen Rasterpunkte abzurufen, mit den notwendigen Korrekturwerten

für die horizontale und vertikale Lage auf dem Bildschirm des auszugebenden Zeichens zu versehen und schliesslich an ein Ausgabeprogramm zu übergeben.

Die Nachteile des Vorgehens lagen in der erschwerten Lesbarkeit der ausgegebenen Schrift, dem geringen Umfang des ausgebbaren Textes (beides begründet durch die geringe Anzahl der zur Verfügung stehenden Rasterpunkte) und vor allem in der langsamen Ausgabegeschwindigkeit, die sich aus der Vielzahl der zur Ausgabe eines einzelnen Zeichens erforderlichen Rasterpunkte ergab.

Die Anschaffung des Datensichtgeräts (Ausgabekapazität: 24 Zeilen zu je 80 Zeichen) ermöglichte dann zwar die schnellere und optisch einwandfreie Ausgabe grösserer Textmengen, erforderte jedoch auch eine mit grossem Zeitaufwand verbundene Umstellung des bis dahin erarbeiteten Systems.

Auch besteht z.B. ein grossrer Teil der Arbeit während der Erprobungsphase eines solchen Systems in der Suche nach oder der Abwehr von Störimpulsen, die entweder eine Veränderung der Versuchsergebnisse oder einen völligen Zusammenbruch des Systems bewirken können.

So kam es z.B. während der Testphase mehrfach vor, dass der Computer nicht mehr auf die Steuerung über die Eingabetastatur reagierte, sondern die Konsolschreibmaschine sinnlose Zeichenfolgen ausgab , obwohl im gesamten Programm keinerlei Schreibbefehl für die Konsolschreibmaschine enthalten war.

Nach längeren Überprüfungen ergab sich, dass über zum Zeitpunkt des Versuchs nicht mit dem Computer verbundene Kabel, die für andere Versuche benötigt wurden, Störimpulse in die für den Versuch benötigten Kabel gelangten. Während aufgrund der technischen Bedingungen der Eingabetastatur nur Eingabewerte zwischen 0 und 15 erzeugt werden konnten, traten bei den Störimpulsen Zahlenwerte weit ausserhalb dieses Bereichs auf. Da der Versuchsfortgang jedoch aufgrund der eingehenden Zahlenwerte gesteuert wurde, wurden durch die Störimpulse Programmverzweigungen ausgelöst, für die keine Verzweigungsadresse definiert war. Zufällig gelangte das Programm in diesen Fällen dann zu Standardroutinen, die die allgemeine Ausgabe über die Konsolschreibmaschine bewirken.

2. "SOFTWARE"

Für die zur Verfügung stehende Rechenanlage liegen zwar die "Compiler" für einige Standardprogrammiersprachen - wie z.B. FORTRAN (FORmula TRANslator), eine der Formelsprache der Mathematik ähnliche Programmiersprache - vor, also Programme zur Umsetzung der Befehle aus der Programmiersprache in eine maschinenausführbare Form, und eine Reihe von Programmpaketen, die vor allem die Berechnung statistischer Werte ermöglichen, doch fehlen die speziellen Routinen für Versuche im Dialog mit dem Computer.

Daher galt es zunächst, ein System von Unterprogrammen zu entwickeln, das eine möglichst allgemeine und schnelle Programmierung der automatisierten Testdurchführung gestattet. Dabei sind wir von folgenden Grundüberlegungen ausgegangen:

- Alle Textabschnitte, die im Lauf der Testdurchführung auf den Bildschirm ausgegeben werden sollen, werden vor dem Beginn der Versuche einmal auf der Magnetplatte abgespeichert und während der einzelnen Versuchsabläufe von dort abgerufen. Mit diesem Vorgehen wird eine schnelle Textausgabe ermöglicht.

- Nach Abschluss der Ausgabe einer "Bildschirmseite" muss jeweils eine Reaktion der Versuchsperson erfolgen, bevor das Steuerprogramm den nächsten Schritt ausführt. Dabei sind nur zwei verschiedene Reaktionsweisen zugelassen:

- Die Versuchsperson zeigt an, dass sie den ausgegebenen Text gelesen hat und nun auf weitere Informationen/Anweisungen wartet; Betätigung der Funktionstaste "Start".

- Die Versuchsperson beantwortet eine im Ausgabetext gestellte Frage. Diese Antworten können jedoch nicht in Klartext sondern nur in codierter Form (z.B. Angabe der Nummer der zutreffenden Mehrfach-Wahl-Antwort) eingegeben werden.

Im folgenden werden zunächst die wesentlichen Unterprogramme[1] für den Dialogverkehr zusammengestellt und in ihrer Funktion kurz beschrieben. Mit dieser Zusammenstellung soll zumindest quantititativ verdeutlicht werden, in wieviel Unterschritte ein kognitiv recht einfacher Prozess - wie Frage und Antwort in codierter Form - zerlegt werden muss, damit er von einem Computer durchgeführt werden kann. Es würde jedoch den zur Verfügung stehenden Raum erheblich überschreiten, wenn wir hier versuchen würden, den Zusammenhang aller Unterprogramme und der in ihnen enthaltenen einzelnen Programmschritte aufzuzeigen, ihre Notwendigkeit zu begründen und auf die zur Entwicklung dieser Schritte erforderlichen Analysen einzugehen.

Danach wollen wir beispielhaft auf zwei Unterprogramme des Systems zur automatisierten Testdurchführung (Textausgabe/ Antworteingabe und die programmierte Einführung in die Tastaturbedienung) eingehen. Dabei wird auch ein Teil der eben angesprochenen Schwierigkeiten deutlich werden.

A. Kurzbeschreibung der benötigten Unterprogramme[2]

Neben den angesprochenen Unterprogrammen zur Textausgabe/ Antworteingabe und zum Erlernen der Tastaturbedienung sind Routinen notwendig, die die Häufigkeit einer Antwortkorrektur registrieren, die Zeit zwischen Textausgabe und Reaktion der Versuchsperson messen, die Dauer des gesamten Versuchs registrieren, bzw. es gestatten, den Versuch nach einer vorgegebenen Zeit abzubrechen. Auch müssen eingegebene Antworten auf ihre inhaltliche Richtigkeit überprüft, die einzelnen Ergebnisse ausgegeben werden u.a.m..

---

1) Soweit nicht anders vermerkt wurden diese Programme von mir für diesen Versuch in FORTRAN erstellt.

2) Die Auflistung der Programme befindet sich im Anhang.

ADRUP:[1] Berechnung von Anfangs- und Endadresse des Speicherbereichs für einen bestimmten Textabschnitt auf der Magnetplatte.
Benötigte Unterprogramme: keine

DAEUP: Textausgabe auf ein optisches Display und Ziffereneingabe über eine seperate Tastatur.
Benötigte Unterprogramme: ADRUP, DTXUP, DZFUP, PENUP, STRUP, TIME, ZFZUP

DGBUP: Abfrage der Digitaleingänge auf einen bestimmten Eingabewertbereich.
Benötigte Unterprogramme: IFDIK

DIGUP: Abfrage der Digitaleingänge auf einen bestimmten Eingabewert.
Benötigte Unterprogramme: IFDIK

DTXUP: Zeilenweise Textausgabe auf ein optisches Display.
Benötigte Unterprogramme: PAGUP

DZFUP: Überprüfung der eingegebenen Antworten auf formale Stimmigkeit.
Benötigte Unterprogramme: DGBUP

IFDIK: Modifizierte ASSEMBLER-Routine des Interface-Herstellers zur Feststellung, ob irgendeine Spannung an den Digitaleingängen anliegt.
Benötigte Unterprogramme: keine

FELUP: Überprüfung von Antworten auf Testaufgaben auf die inhaltliche Richtigkeit.
Benötigte Unterprogramme: keine

---

[1] Unterprogramme können - innerhalb gewisser Grenzen - willkürlich benannt werden. Ihre Namen sind hier nur aufgeführt, um den Bezug zwischen den Programmen aufzuzeigen.

PAGUP: Berechnung des Zeilenvorschubs bei der zeilenweisen Textausgabe auf ein optisches Display.
Benötigte Unterprogramme: keine

PENUP: Einzelausgabe der über eine externe Tastatur eingegebenen einzelnen Ziffern einer natürlichen Zahl.
Benötigte Unterprogramme: ZFZUP

PRNUP: Ausgabe der Versuchsdaten über Kartenstanzer und Zeilendrucker.
Benötigte Unterprogramme: keine

RZEIT: ASSEMBLER-Routine von D. LANGHEINRICH zur Messung der gesamten Versuchsdauer.
Benötigte Unterprogramme: keine

STRUP: Textausgabe auf ein optisches Display und Abfragen der "Start"-Taste.
Benötigte Unterprogramme: ADRUP, DIGUP, DTXUP

TASUP: Programmierte Unterweisung zum Erlernen der Tastaturbedienung.
Benötigte Unterprogramme: DAEUP, STRUP

TIME: ASSEMBLER-Routine des Interface-Herstellers zum Starten und Stoppen des Zeitregisters, sowie zur Abfrage von einzelnen und kumulierten Zeiten.
Benötigte Unterprogramme: keine

TSTUP: Gesamtdurchführung eines diagnostischen Testverfahrens.
Benötigte Unterprogramme: DAEUP, RZEIT, TIME

ZFZUP: Berechnung einer natürlichen Zahl aufgrund der eingegebenen Ziffern.
Benötigte Unterprogramme: keine

ZTVGL: Kontrolle der Übereinstimmung der Zeittakte in beiden Zeitregistern.
Benötigte Unterprogramme: RZEIT, TIME

Wie sich meist schon an Anzahl und Art der von einem Unterprogramm aufgerufenen weiteren Unterprogramme ablesen lässt, sind die einzelnen Routinen in ihrem Hirarchiegrad recht unterschiedlich.

Betrachtet man dieses System als Modell eines Dialogs, so finden sich Routinen, die den entsprechenden menschlichen Handlungen sehr ähnlich sind (z.B. IFDIK: das aufmerksame Warten auf die Antwort des Gesprächspartners; FELUP: Überprüfen der Richtigkeit der Antwort auf eine Testaufgabe), und solche, die sich aus den technischen Anforderungen des Systems ergeben (z.B. PAGUP: Berechnung des Zeilenvorschubs für die Textausgabe).

Insgesamt ist mit diesem System von Unterprogrammen die Möglichkeit gegeben, einen beliebigen Test in Mehrfach-Wahl-Form in relativ kurzer Zeit für die computergesteuerte Darbietung aufzubereiten.

B. Textausgabe und Antworteingabe (Unterprogramm DAEUP)

Wir wollen nun am Beispiel des Unterprogramms DAEUP, das die Bildschirmausgabe eines Textes und die Eingabe der Versuchspersonen-Antwort über eine externe Tastatur überwacht und steuert, einen gewissen Überblick über die Vielzahl der zu berücksichtigenden Schritte bei der Entwicklung eines automatisierten Systems geben.- vgl. dazu auch Abb. 2.

Vor dem Beginn der eigentlichen Bildschirmausgabe sind zunächst eine Reihe von Speichern (wie. z.B. zur Registrierung der von der Versuchsperson vorgenommenen Antwortkorrekturen) auf den Wert Null zu setzen.

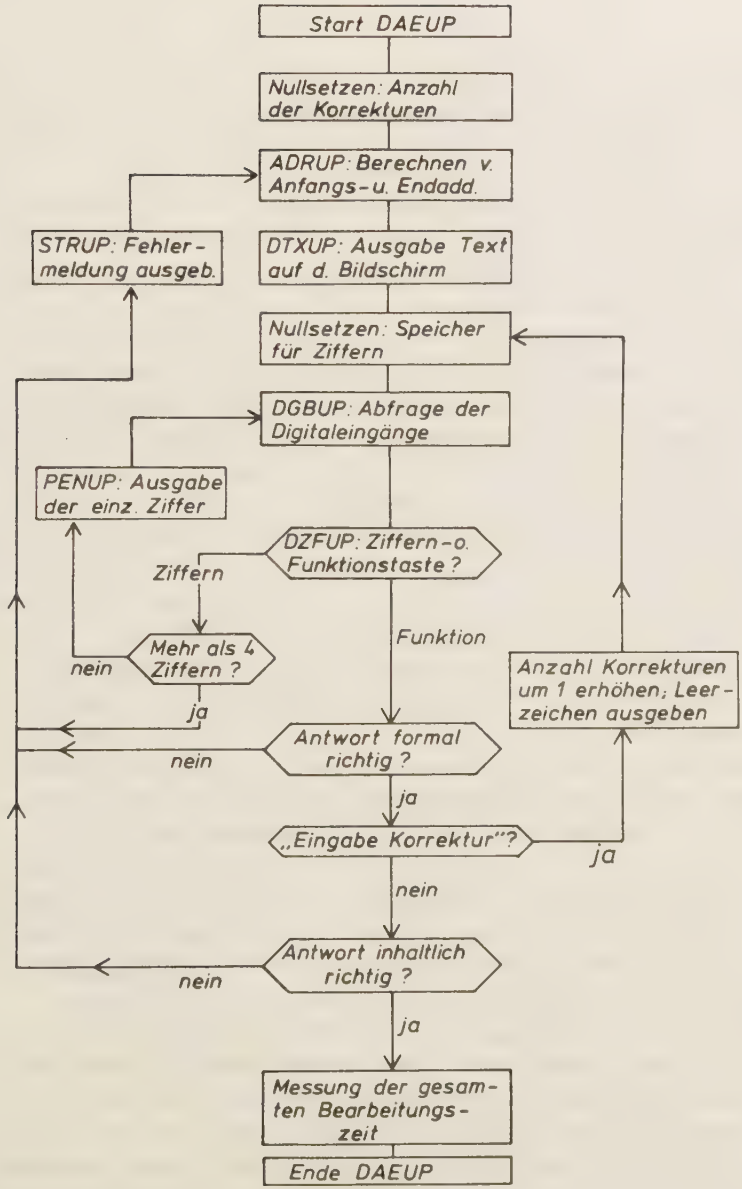


Abb. 2: Vereinfachtes Flussdiagramm für das Unterprogramm DAEUP - Erläuterungen im Text.

Im Unterprogramm ADRUP werden dann Anfangs- und Endadresse des zeilenweise auf der Magnetplatte gespeicherten Ausgabetextes berechnet. Die eigentliche Textausgabe geschieht im Unterprogramm DTXUP, indem jede Textzeile von der Platte gelesen, ihr Abstand zur ersten Bildschirmzeile berechnet und der Text dann ausgeschrieben wird.

Nach Abschluss der Ausgabe einer "Bildschirm-Seite" muss nun eine Reaktion der Versuchsperson erfolgen, bevor das Steuerprogramm weitergeht. Zunächst wird der Speicher, der die von der Versuchsperson eingebenen Ziffern aufnehmen soll, auf den Wert Null gesetzt. Danach werden über das Unterprogramm DGBUP so lange die Digitaleingänge des Interface abgefragt, bis eine dort auftretende Spannung die Betätigung irgendeiner Taste durch die Versuchsperson anzeigt. Handelt es sich hier um eine Zifferntaste, so ist zu überprüfen, ob die bisher insgesamt eingegebene Zahl nun mehr als vier Stellen umfasst. In diesem Fall muss eine entsprechende Fehlermeldung an den Probanden ausgegeben werden. Liegt die Zahl der eingegebenen Ziffern unter fünf, wird die zuletzt eingetastete Ziffer zur Kontrolle durch die Versuchsperson auf den Bildschirm augegeben, und das Unterprogramm DGBUP wartet auf weitere Reaktionen der Versuchsperson.

Hat die Versuchsperson eine Funktionstaste betätigt, wird zunächst geprüft, ob diese Reaktion formal richtig ist (richtig ist "Eingabe Ende" oder "Eingabe Korrektur" nach vorheriger Zahleneingabe; falsch ist, ausser der Betätigung dieser beiden Tasten vor der Zahleneingabe, auch die Betätigung der "Start-Taste"). Bei einem formalen Fehler wird wiederum eine enstprechende Fehlermeldung ausgegeben, und dann der zu beantwortende Text beginnend mit der Routine ADRUP erneut auf den Bildschirm geschrieben. Wurde von der Versuchsperson formal richtig die Korrekturtaste gedrückt, so ist das Register für die Registrierung der Korrekturen um 1 zu erhöhen, die bisher auf dem Bildschirm ausgegebene Zahl mit Leerzeichen zu überschreiben und, nachdem der Speicher

für die von der Versuchsperson eingegebenen Ziffern den Wert Null erhalten hat, erneut zum Unterprogramm DGBUP zu verzweigen.

Handelt es sich bei formal richtiger Betätigung der Funktionstasten nicht um eine Antwortkorrektur, so ist die Taste "Eingabe-Ende" gedrückt worden. Dann muss überprüft werden, ob die eingegebene Antwort der vorgegebenen Frage entspricht und damit weiter verarbeitet werden kann. So ist es z.B. möglich, auf eine Frage mit 5 Antwortmöglichkeiten formal richtig die Zahl 6 einzugeben, doch muss dann eine entsprechende Fehlermeldung an den Probanden ausgegeben werden, da die Bedeutung dieser Antwort nicht definiert ist. Ist die Antwort in diesem Sinn richtig, so wird die gesamte Bearbeitungszeit gemessen, und es kann z.B. eine andere Frage ausgegeben werden.

C. Die programmierte Einführung in die Bedienung der Eingabetastatur (Unterprogramm TASUP)

Vor dem eigentlichen Beginn einer automatisierten Testdurchführung muss den Probanden Gelegenheit gegeben werden, sich mit dem äusseren Versuchsaufbau vertraut zu machen und eine gewisse Übung im Umgang mit den eingeschränkten Kommunikationsmöglichkeiten eines automatisierten Systems zu erhalten. Dies geschieht bei dem hier beschriebenen System durch das Unterprogramm TASUP - vgl. dazu auch Abb. 3.

Auf den Bildschirm wird der Text 1 ausgegeben[1)], in dem der allgemeine Versuchsaufbau beschrieben wird. Danach wartet das Steuerprogramm so lange, bis die Start-Taste betätigt wird (die Betätigung der übrigen Tasten wird von dem Programm ignoriert). Es folgt dann die Ausgabe von Text 2 mit einer genaueren Erklärung der Funktion der Start-Taste.

---

1) Um Platz zu sparen, wird hier der Inhalt der einzelnen Texte nur kurz angegeben. Die vollständigen Texte enthält die Abb. 4.

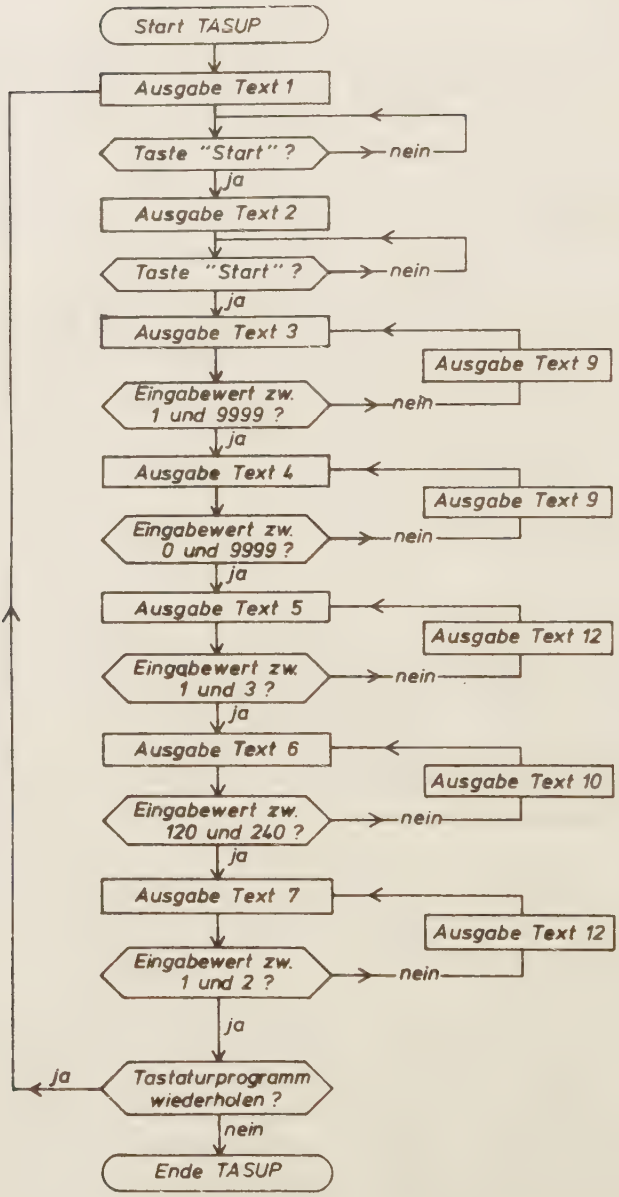


Abb. 3: Vereinfachtes Flussdiagramm für das Unterprogramm TASUP - Erläuterungen im Text.

Text 1

SEHR VEREHRTE 'VERSUCHSPERSON',

SIE WERDEN NUN DEN DIALOG MIT DER DATENVERARBEITUNGSANLAGE 'IBM 1130' AUFNEHMEN. DA IHR 'GESPRAECHSPARTNER' EINE MASCHINE IST, SIND DIE MOEGLICHKEITEN ZUM DIALOG LEIDER RECHT BESCHRAENKT UND DARUEBER HINAUS STARREN KOMMUNIKATIONSREGELN UNTERWORFEN.

ZUNAECHST GEHT ES NUR DARUM, DASS SIE SICH SCHRITTWEISE MIT DIESEN KOMMUNIKATIONSREGELN VERTRAUT MACHEN UND IHRE ANWENDUNG EINUEBEN. - DIESER LERNPROZESS WIRD BEREITS DURCH DEN COMPUTER GESTEUERT.

* SIE ERHALTEN INFORMATIONEN VOM COMPUTER NUR UEBER DEN VOR IHNEN STEHENDEN BILDSCHIRM.
* INFORMATIONEN FUER DEN COMPUTER KOENNEN SIE NUR UEBER DIE VOR IHNEN STEHENDE TASTATUR EINGEBEN. - DIESE TASTATUR ENTHAELT IHREM VERWENDUNGSZWECK NACH ZWEI VERSCHIEDENE ARTEN VON TASTEN.

  - 10 TASTEN FUER ZIFFERN (1,2,3,4,5,6,7,8,9,0)
  - 3 FUNKTIONSTASTEN (START - EINGABE ENDE - EINGABE KORREKTUR)
    (DIE BEIDEN UEBRIGEN TASTEN WERDEN NICHT BENOETIGT)

WENN SIE NUN 'START' DRUECKEN, ERHALTEN SIE DIE NAECHSTE INFORMATION.

Text 2

ALLGEMEIN WIRD DIE TASTE 'START' ALSO DAZU BENUTZT, WEITERE INFORMATIONEN VOM COMPUTER ZU ERHALTEN, OHNE DASS VON IHNEN EINE BESTIMMTE ANTWORT (Z.B. DIE EINGABE EINER ZAHL) ERWARTET WIRD.

DURCH BETAETIGEN DER TASTE 'START' ZEIGEN SIE AN, DASS SIE DEN TEXT AUF DEM BILDSCHIRM GELESEN HABEN UND NUN EINE NEUE INFORMATION ERWARTEN.

BILDSCHIRMTEXTE DER HIER BESCHRIEBENEN ART ERKENNEN SIE DARAN, DASS IN DER ZULETZT AUSGEGEBENEN BILDZEILE DIE FOLGE /START/ ERSCHEINT.
DEM AUFBAU DER TASTATUR, MIT DER ALLEIN SIE MIT DEM COMPUTER IN VERBINDUNG TRETEN KOENNEN, HABEN SIE SICHER SCHON ENTNOMMEN, DASS SIE KEINEN BELIEBIGEN TEXT ALS ANTWORT EINGEBEN, SONDERN NUR IN FORM VON ZAHLEN ANTWORTEN KOENNEN.

Text 3

FUER DIE EINGABE VON ZAHLEN GILT

* ES KOENNEN NATUERLICHE ZAHLEN MIT BIS ZU VIER STELLEN UND DIE '0' (NULL) EINGETASTET WERDEN (EINGABEBEREICH= 0,1,2 ... 9997,9998,9999).
* DIE ZIFFERN EINER ZAHL WERDEN VON LINKS NACH RECHTS - ALSO BEGINNEND MIT DER ZIFFER, DIE DEN HOECHSTEN STELLENWERT HAT - EINGETASTET.
* 'FUEHRENDE NULLEN' BRAUCHEN SIE BEI DER EINGABE NICHT ZU BERUECKSICHTIGEN. SIE TASTEN ALSO Z.B. '85' UND NICHT '0085' EIN.

EINE VON IHNEN ALS ANTWORT EINGETASTETE ZAHL KANN ALSO BIS ZU VIER ZIFFERN UMFASSEN. DER COMPUTER KANN NATUERLICH NICHT IMMER VORHERSEHEN, WIEVIELE ZIFFERN IHRE 'ANTWORT-ZAHL' AUF EINE BESTIMMTE FRAGE UMFASST. DAHER MUESSEN SIE JEWEILS ANGEBEN, WANN SIE ALLE ZIFFERN EINER ZAHL EINGETASTET HABEN. DIES GESCHIEHT, INDEM SIE DIE TASTE 'EINGABE ENDE' BETAETIGEN.

GEBEN SIE NUN BITTE EINE BELIEBIGE ZAHL ZWISCHEN 1 UND 9999 EIN UND DRUECKEN SIE DANN DIE TASTE 'EINGABE ENDE'.

Abb. 4: Zusammenstellung der im Unterprogramm TASUP verwendeten Ausgabe-Texte.

Text 4

NACHDEM SIE EINE ZAHL ALS ANTWORT EINGETASTET UND 'EINGABE ENDE' BETAETIGT HABEN, GEHT DER COMPUTER - SO WIE EBEN GESCHEHEN - ZUM JEWEILS NAECHSTEN PROGRAMMTEIL WEITER.

WIE SIE VIELLEICHT BEIM LETZTEN PROGRAMMSCHRITT GESEHEN HABEN, WURDE DIE VON IHNEN EINGETASTETE ZAHL AM ENDE DES BILDSCHIRMTEXTES UNTER DER FOLGE
=== ANTWORT === AUSGEGEBEN. DAZU FOLGENDE ANMERKUNGEN

- JEDER TEXT, DER EINE FRAGE AUFWIRFT, DIE SIE DURCH EINGABE EINER BESTIMMTEN ZAHL BEANTWORTEN, ENTHAELT IN DER LETZTEN AUSGABEZEILE DIE FOLGE
  === ANTWORT === .
- VON IHNEN EINGETASTETE ZIFFERN WERDEN IMMER RECHTS UNTEN AUF DEM BILDSCHIRM UNTERHALB DER FOLGE === ANTWORT === AUSGEGEBEN.

WENN SIE SICH EINMAL BEI DER ZIFFERNEINGABE VERTIPPEN, KOENNEN SIE DIESEN FEHLER DURCH BETAETIGEN DER TASTE 'EINGABE KORREKTUR' VERBESSERN. NACH DRUECKEN VON 'EINGABE KORREKTUR' WERDEN DIE AUF DEM BILDSCHIRM AUSGEGEBENEN ZIFFERN GELOESCHT, UND SIE KOENNEN DANN DIE RICHTIGE ANTWORT EINGEBEN.

GEBEN SIE NUN BITTE EINE BIS ZU VIERSTELLIGE ZAHL EIN, LOESCHEN SIE DIESE MIT 'EINGABE KORREKTUR' UND GEBEN SIE DANN DIE 'RICHTIGE' ZAHL EIN. (VERGESSE SIE BITTE NICHT NACH DER 'RICHTIGEN' EINGABE DIE TASTE 'EINGABE ENDE' ZU BETAETIGEN.)

Text 5

JE NACH FORM DER FRAGE LASSEN SICH ZWEI VERSCHIEDENE ANTWORTARTEN UNTERSCHEIDEN
1. DIE 'MEHRFACH-WAHL-ANTWORT'
2. DIE 'FREIE ANTWORT'

1. DIE 'MEHRFACH-WAHL-ANTWORT'
****

ZU EINER FRAGE WERDEN VERSCHIEDENE ANTWORTMOEGLICHKEITEN VORGEGEBEN. SIE ANTWORTEN, INDEM SIE DIE NUMMER EINER DER MOEGLICHKEITEN, NAEMLICH DER ZUTREFFENDEN, EINTASTEN. - DAZU EIN BEISPIEL

- SIND SIE VON DER KOERPERGROESSE HER EHER

  (1) KLEIN
  (2) MITTEL
  (3) GROSS

Text 6

2. DIE 'FREIE ANTWORT'
****

AUS DER FRAGESTELLUNG ERGIBT SICH, DASS NUR EINE, NAEMLICH DIE IHNEN ZUTREFFEND ERSCHEINENDE ZAHL ZUR BEANTWORTUNG IN FRAGE KOMMT. DIESE ZAHL TASTEN SIE DANN EIN. - EIN BEISPIEL

- WIE GROSS SIND SIE (IN ZENTIMETERN)

Abb. 4: (Fortsetzung)

Text 7

SIE KENNNEN JETZT DIE FUER DEN DIALOG MIT DER IBM 1130 NOTWENDIGEN KOMMUNIKATIONSREGELN.

* WOLLEN SIE NUN

(1) MIT DER EIGENTLICHEN AUFGABE BEGINNEN

(2) DIE BESCHREIBUNG DER EINZELNEN TASTEN UND IHRER BEDEUTUNG NOCH EINMAL WIEDERHOLEN.

Text 9

DIE VON IHNEN EINGETASTETE ZAHL UMFASST MEHR ALS VIER ZIFFERN UND MUSSTE DAHER GELOESCHT WERDEN.

NACH BETAETIGUNG DER TASTE START GEHT DAS PROGRAMM ZUM LETZTEN SCHRITT ZURUECK, UND SIE KOENNEN DANN IHRE ANTWORT ERNEUT EINGEBEN.

Text 10

DIE VON IHNEN ANGEGEBENE KOERPERGROESSE LIEGT SO WEIT AUSSERHALB DES NORMALEN BEREICHS, DASS EIN EINGABEFEHLER VERMUTET WERDEN MUSS.

DRUECKEN SIE BITTE DIE TASTE 'START' UND GEBEN SIE DANN IHRE ANTWORT ERNEUT EIN.

Text 12

DIE VON IHNEN EINGEGEBENE ZAHL ENTSPRICHT KEINER DER VORGEGEBENEN KATEGORIEN EINER MEHRFACH-WAHL-ANTWORT.

BETAETIGEN SIE BITTE DIE TASTE 'START' UND GEBEN SIE DANN IHRE ANTWORT ERNEUT EIN.

Abb. 4: (Fortsetzung)

Nachdem die Versuchsperson die Taste "Start" gedrückt hat, erhält sie im Text 3 eine Beschreibung des möglichen Bereiches bei der Eingabe von Zahlen und der Bedeutung der "Eingabe-Ende"-Taste; auch wird sie zur Eingabe einer beliebigen Zahl innerhalb dieses Bereiches aufgefordert. - Diese und die weiteren Schritte des Tastatur-Programms werden jeweils über das oben beschriebene Unterprogramm DAEUP gesteuert.

Der Text 4 erläutert die Funktion der Taste "Eingabe-Korrektur"; auch wird nochmals an "Eingabe-Ende" erinnert.

Mit Text 5 werden die beiden unterschiedlichen Möglichkeiten bei der Zahleneingabe ("Mehrfach-Wahl-Antwort" und "freie Antwort") eingeführt; eine Beispielaufgabe für die "Mehrfach-Wahl-Antwort" ist zu beantworten. Bei einer Antwort ausserhalb der vorgegebenen Wahlmöglichkeiten erhält die Versuchsperson eine entsprechende Fehlermeldung (Text 12).

Im Text 6 wird ein Beispiel für die "freie Antwort" gegeben. Auch hier wird der Versuchsperson bei einer falschen Antwort mit der Ausgabe von Text 10 deutlich gemacht, dass ihre Antworten durch den Computer überprüft werden.

Bei Text 7 wird der Versuchsperson die Möglichkeit eingeräumt, zwischen der Wiederholung des Tastatur-Programms und dem Beginn der Versuche zu wählen. Im ersten Fall beginnt das Programm wiederum mit Text 1, im zweiten wird zu der eigentlichen Testdurchführung verzweigt.

Neben den oben angesprochenen Zielen der programmierten Unterweisung zur Einführung in die Tastaturbedienung (Vertrautwerden mit dem äusseren Versuchsaufbau, Erlernen der Kommunikationsregeln) sollen den Versuchspersonen in der Übungsphase vor allem noch zwei Punkte deutlich werden:

- Das automatisierte System erzwingt von der Versuchsperson eine Antwort. Denn so lange keine Antwort eingegeben wird, bleibt das Programm bei dem entsprechenden Schritt stehen.

- Bei der Beantwortung lässt das System kein Ausweichen zu. Liegt die Antwort ausserhalb des vorgegebenen Bereichs, so erfolgt immer eine entsprechende Fehlermeldung, und es wird eine neue Antwort gefordert.

III. NOTWENDIGE VORAUSSETZUNGEN FÜR DIE ÜBEREINSTIMMUNG VON TESTERGEBNISSEN AUS HERKÖMMLICHER UND AUTOMATISIERTER TESTDURCHFÜHRUNG (1. VERSUCHSSERIE)

Mit der oben beschriebenen Hard- und Software haben wir die äusseren Voraussetzungen für eine automatisierte Testdurchführung geschaffen. Doch ergibt sich damit die Frage, auf welche Art ein bestehender Test am besten in eine automatisierte Form übertragen werden kann. In entsprechenden Untersuchungen (z.B. ELWOOD, 1973; JOHNSON & MIHAL, 1973; DUNN, LUSHENNE & O'NELL, 1972) finden sich keine weiterführenden Hinweise, sondern die Autoren scheinen implizit davon auszugehen, dass die von ihnen gewählte Form der Übertragung - unter Berücksichtigung der jeweiligen technischen Bedingungen - die richtige ist.

1. ENTWICKLUNG DER UNTERSUCHUNGSFRAGE

In diesem Teil der Untersuchung haben wir also zunächst zu überprüfen, wieweit bei der Automatisierung eines Testverfahrens die Art der Testdurchführung an die herkömmliche Testdarbietung anzugleichen ist, damit die Ergebnisse aus beiden Durchführungsformen miteinander vergleichbar werden. Dabei beschränken sich diese Überlegungen beispielhaft auf Leistungstests in der gebräuchlichsten Form der Testdurchführung, auf die Gruppe der "Papier- und - Bleistift" - Verfahren, und beziehen sich dort auf Aufgaben vom Typ der "Mehrfach-Wahlantwort."

A. Herkömmliche und automatisierte Testdurchführung

Ausgehend von einer kurzen Beschreibung der herkömmlichen Testdurchführung wollen wir verschiedene Möglichkeiten der automatisierten Aufgabendarbietung erörtern -vgl. auch GLAUBITZ, 1975. Es ist vorauszuschicken, dass bei der Analyse der üblichen Test-

situation sicher eine Vielzahl spezifischer Kennzeichen herausgestellt werden könnten; doch sind die "Simulierungsmöglichkeiten" dieser Situation durch den Computer so gering, dass wir uns auf die wichtigsten Faktoren beschränken wollen.

a. Herkömmliche Testdurchführung

Die herkömmliche Testdurchführung enthält eine Reihe von Voraussetzungen, die wahrscheinlich aus Gewohnheit stillschweigend angenommen werden. Jedoch sind sie bei dem Versuch, einen solchen Test in eine automatisierte Form zu übertragen, zu berücksichtigen und damit zunächst herauszuarbeiten.

Wenn der Proband den Untersuchungsraum betritt, findet er dort meist das Testheft vor, oder es wird ihm nach einer allgemeinen Einführung ausgehändigt. Dabei besteht fast immer die Möglichkeit, das Testheft kurz durchzublättern und sich so einen ersten Eindurck von Art und Menge der zu bearbeitenden Aufgaben zu verschaffen.
Wenn auch der Proband während der eigentlichen Testdurchführung ohne Kontakt mit dem Untersuchungsleiter arbeitet, hat er in der Instruktionsphase die Möglichkeit, Verständnisfragen zu stellen.

Bei der Testdurchführung besteht aufgrund der Anlage der Testhefte für den Probanden die Möglichkeit,

- o sich ständig einen Überblick über bereits geleistete Arbeit und die Anzahl der noch zu bearbeitenden Items zu verschaffen;
- o einzelne, im Augenblick zu schwere Aufgaben zunächst auszulassen und eventuell zu einem späteren Zeitpunkt noch einmal in Angriff zu nehmen;
- o Fehler sofort beim Ankreuzen zu verbessern;
- o bei späterem Durchlesen Fehler zu verbessern, die etwa zunächst übersehen worden sind, oder die erst deutlich werden, nachdem aufgrund der Bearbeitung einer Reihe von Aufgaben der Lösungsweg gefunden ist.

b. Verschiedene Möglichkeiten der automatisierten Testdarbietung

aa. Das Konzept einer einfachen automatisierten Darbietung

Stellen wir uns im Gegensatz dazu eine möglichst einfache Form einer automatisierten Darbietung vor:

Der Proband wird vom Testleiter in den Untersuchungsraum geführt und zunächst mit dem technischen Aufbau vertraut gemacht. Es wird ihm erklärt, dass Anweisungen und die eigentlichen Testaufgaben auf dem Bildschirm erscheinen und über Knopfdruck zu beantworten sind. Danach wird er mit dem automatischen System allein gelassen. Der Proband wird nun die Anweisung auf dem Bildschirm lesen, bis er glaubt, sie verstanden zu haben; nach Betätigung einer entsprechenden Taste wird ihm die erste Aufgabe vorgegeben, die er durch Tastendruck beantwortet. Sofort erscheint eine neue Aufgabe.

Wir sehen, dass diese Form der automatisierten Testdurchführung sich wesentlich von der herkömmlichen Art unterscheidet; so kann der Proband, da die eingetastete Zahl nicht auf dem Bildschirm erscheint, nicht einmal kontrollieren, ob er auch wirklich die entsprechende Taste gedrückt hat. Damit entfällt auch die Möglichkeit, etwaige Eingabefehler zu verbessern. Da jede Antworteingabe aber sofort die Ausgabe einer neuen Testaufgabe bewirkt, kann die Versuchsperson zusätzlich verunsichert werden. Deswegen sollten ihr zumindestens die Möglichkeiten zur Kontrolle der eingetasteten Werte und zur nachträglichen Korrektur dieser Werte gegeben werden.

bb. Automatisierte Testdarbietung mit Eingabekorrektur

Überträgt man die Begriffe Eingabekontrolle und Eingabekorrektur auf die "Papier-und-Bleistift-Form" der Testdurchführung, so wird damit nur der eigentlich selbstverständliche Sachverhalt beschrieben, dass die Versuchspersonen sieht, welche Antwortmöglichkeit sie markiert hat, und dass sie im Anschluss daran die angekreuzte Antwort verbessern kann.

Wir wollen an dieser Stelle noch einmal auf das Flussdiagramm des Unterprogramms DAEUP (Abb. 2) verweisen. Sieht man von den dort zusätzlich eingeführten Programmschritten zur Zeitmessung ab, so verwirklicht dieses Programm gerade die eben angesprochene Eingabekontrolle und Eingabekorrektur mit dem zusätzlichen Faktor, dass das Programm Auskunft darüber geben kann, wie oft der Proband bei einem Item "radiert" hat.
Dieser Rückverweis scheint uns aus zwei Gründen notwendig zu sein: zum einen kann gezeigt werden, wie aufwendig es ist, solch einfache Handlung wie das Markieren einer Antwortmöglichkeit durch ein automatisches System entsprechend darzustellen. Zum anderen wird hier schon deutlich, dass die Probanden mit der Maschine in einer Art kommunizieren, deren Regeln erst erlernt werden müssen; zusätzlich werden dabei Handlungsabläufe, über die die Versuchsperson sonst überhaupt nicht nachzudenken braucht, direkt angesprochen und damit eventuell problematisiert. Um im Bild zu bleiben: Bei der herkömmlichen Testdurchführung wird niemand die Versuchsperson daran erinnern, dass es unsinnig ist, eine Antwortmarkierung auszuradieren, bevor diese überhaupt markiert worden ist.

Auch mit den Möglichkeiten der Eingabekontrolle und Eingabekorrektur ist die automatisierte Form nicht restlos an den tatsächlichen Verlauf einer herkömmlichen Testdurchführung angeglichen; zwei Gründe sprechen gegen eine vollständig Übertragung. Einerseits wird der Programmieraufwand beim Zerlegen der Testsituation in immer kleinere Einzelschritte unverhältnismässig gross.
Vor allem aber werden mit komplizierteren Programmen auch die von der Versuchsperson zu erlernenden Kommunikationsregeln immer komplexer. Dann wäre sicher bald ein Punkt erreicht, wo mit einem automatisierten Verfahren nicht mehr der Inhalt des Tests sondern nur noch das Instruktionsverständnis überprüft wird.
Bei einer praktikablen Form muss also dem Probanden ermöglicht werden, sich wie gewohnt zu verhalten, doch dürfen damit die Verhaltensregeln selbst nicht zu kompliziert werden.

cc. Einzel- und Gesamtausgabe der Testaufgaben

Wir sind bisher stillschweigend davon ausgegangen, dass bei der automatisierten Durchführung jeweils nur ein Item auf dem Bildschirm erscheint, Zwar ist die Kapazität des Bildschirms geringer als die einer Druckseite, doch liessen sich zumindest mehrere Aufgaben gleichzeitig ausgeben.

Eine zusätzliche Kapazitätserweiterung bietet der sogenannte "roll-mode"; bei vollgeschriebenem Bildschirm wird jeweils die oberste Zeile gelöscht, der weitere Text wird jeweils um eine Zeile höher geschoben, während die unterste Bildzeile frei wird. Doch spricht allein schon die Unruhe, die die ständigen Verschiebungen mit sich bringen, gegen diese Darbietungsform.

Die gleichzeitige Darbietung mehrerer Items wirft zusätzlich das Problem auf, die einzelnen Antworten den Items eindeutig zuzuordnen. Bei starrer Reihenfolge könnte jeweils durch einen Pfeil optisch angezeigt werden, welche Aufgabe gerade zur Beantwortung ansteht - Eingabekontrolle und Eingabekorrektur würden wie oben beschrieben vor sich gehen - doch wäre damit dem Probanden noch nicht die Möglichkeit eingeräumt, im nachherein als falsch erkannte Lösungen zu korrigieren. Wollte man ihm aber dies bei gleichzeitiger Darbietung mehrerer Aufgaben einräumen, so würde dann wiederum ein sehr komplexes Programm- und Regelsystem notwendig.

dd. Zusätzliche Faktoren

Wenn auch vieles für die Einzeldarbietung der Testaufgaben spricht, so fehlt bei dieser Darbietungsart dem Probanden jedoch die Kontrollmöglichkeit über die bereits geleistete und noch zu leistende Arbeit. Hier lässt sich relativ einfach Abhilfe schaffen, indem vor Beginn jedes Tests mitgeteilt wird, wieviele Aufgaben der Test enthält, und indem die einzelnen Aufgaben durchnummeriert werden.

Schwieriger wird es, dem Probanden die Möglichkeit einzuräumen, einzelne Aufgaben zu überspringen. Beginnen wir hier mit dem einfachsten Fall: wir können in der multiple-choice-Form bei jedem Item eine zusätzliche Antwortmöglichkeit "Item soll ausgelassen werden" einführen. Doch stellt sich dann die Frage, wann die übersprungenen Aufgaben bearbeitet werden sollen.
Man könnte hier daran denken, nach Bearbeitung des gesamten Tests die zunächst zurückgestellten Items in der Standardreihenfolge erneut vorzugeben. Gegenüber der herkömmlichen Testdurchführung, bei der der Proband in der Reihenfolge der Aufgabenbearbeitung jederzeit vor- und zurückspringen kann, bedeutet dies sicher eine recht willkürlich vorgenommene Einschränkung.
Auf der anderen Seite lässt sich ein variableres Programm kaum in der Praxis anwenden; es wäre dann z.B. bei jeder Aufgabe eine zweite zusätzliche Antwortmöglichkeit ("vorher übersprungenes Item soll jetzt beantwortet werden") einzuführen, worauf in einem zusätzlichen Schritt zunächst gefragt werden müsste, welches vorher ausgelassene Item zu diesem Zeitpunkt beantwortet werden soll. Das aber würde wiederum bedeuten, dass über den Bildschirm zunächst eine Liste der von der Versuchsperson bisher beantworteten bzw. nicht beantworteten Aufgaben auszugeben wäre, aus der dann die zur Bearbeitung gewünschte Aufgabe ausgewählt werden müsste.

Wir wollen diese Überlegungen hier nicht weiter ausführen, da sicher deutlich geworden ist, welch komplexes Programm benötigt werden würde, um einen bestehenden Test in eine möglichst entsprechende automatisierte Form zu übertragen. Das bedeutet aber für Erstellung automatisierter Testverfahren, dass der computergesteuerte Test gegenüber der herkömmlichen Durchführung immer einer Reihe von Beschränkungen unterworfen ist.

B. Ziel der 1. Versuchsserie

Bevor wir daran gehen können, den diagnostischen Gewinn einer automatisierten Testdurchführung zu untersuchen, der über den Wert der bei herkömmlicher Testführung anfallenden diagnostischen Informationen hinausgeht, muss sichergestellt sein, dass der Effekt einer computergesteuerten Durchführung zumindest dem der herkömmlichen entspricht. Dazu sind etwa folgende Fragen abzuklären:

- Gelingt es den Probanden, sich auf das oben beschriebene System zur automatisierten Testdurchführung einzustellen?

- Wieweit muss die Art der automatisierten Aufgabendarbietung an die herkömmliche Testdurchführung angenähert werden, damit die Ergebnisse aus beiden Darbietungsarten vergleichbar bleiben?

- Werden die beiden Durchführungsarten von den Probanden als unterschiedlich belastend erlebt, und hat das Auswirkung auf die Testleistung?

Die erste Untersuchungsfrage, ob es den Probanden gelingt, sich auf die automatisierte Testdurchführung einzustellen, lässt sich nur schwer in eine statistisch überprüfbare Hypothese fassen. Daher wollen wir davon ausgehen, dass das in Vorversuchen erprobte (vgl. GLAUBITZ, 1975) und aufgrund der dabei gewonnenen Erfahrungen modifizierte Testsystem den Probanden keine Schwierigkeiten bereitet. Diese Annahme ist dann zu verwerfen, wenn es den Probanden nicht gelingt, sich auf diese Art der Versuchsdurchführung nach kurzer Zeit einzustellen, bzw. sie den Versuch sogar abbrechen.

Für die Art der Aufgabendarbietung gehen wir von der Annahme aus, dass mit zunehmender Angleichung der automatisierten Form an die herkömmliche Durchführung

die Übereinstimmung zwischen den Ergebnissen aus beiden Durchführungsarten grösser wird.

Wie oben dargestellt sind für die herkömmliche Durchführung eines "Papier-und-Bleistift-Tests" zwei Faktoren besonders kennzeichnend:

- o Der ständige Überblick über die geleistete und noch zu leistende Arbeit.
- o Die Möglichkeit, die Bearbeitung einzelner Items zunächst zurückzustellen.

Damit bieten sich zur Überprüfung der Annäherung eines automatisierten Verfahrens an die herkömmliche Durchführung vier Kombinationen an, bei denen diese Faktoren vorhanden (+) oder nicht vorhanden (-) sind:

Faktor	Kombination 1	2	3	4
"Überblick":	-	+	-	+
"Zurückstellen":	-	-	+	+

Die 3. Kombination, bei der dem Probanden zwar die Möglichkeit eingeräumt wird, einzelne Items zurückzustellen, er aber auf der anderen Seite nicht weiss, wieviele Aufgaben insgesamt zu bearbeiten sind, ist unter der hier zu untersuchenden Frage nicht sinnvoll. Denn damit käme eine weitere Variable dazu, die man etwa als "Risikobereitschaft" des Probanden bezeichnen könnte.

Wir setzen nun voraus, dass die Kombinationen

"Überblick" (-) / "Zurückstellen" (-)
"Überblick" (+) / "Zurückstellen" (-)
"Überblick" (+) / "Zurückstellen" (+)

eine immer stärkere Annäherung der automatisierten an

die herkömmliche Testdurchführung bedeuten und erwarten, dass mit zunehmender Annäherung die Übereinstimmung zwischen den Ergebnissen aus beiden Durchführungsarten immer grösser wird.

Inbezug auf die Auswirkungen der beiden Durchführungsarten auf die Probanden wird erwartet:

- Beide Situationen werden insgesamt als gleich belastend bzw. nicht-belastend erlebt.
- Bei einer Unterscheidung der Probanden jeweils nach der bevorzugten Durchführungsart werden in dieser die besseren Ergebnisse erzielt. (vgl. CAHOON, PETERSON & WATSON, 1968)

C. Versuchsaufbau und Versuchsdurchführung[1)]

Die Stichprobe umfasste 50 Studenten der hiesigen Universität, die zum grössten Teil im Hauptfach Psychologie studieren (32 weiblich, 18 männlich; Alter zwischen 19 und 44 Jahren; Durchschnittsalter 24 Jahre). Sie wurden über einen Aushang angeworben, in dem lediglich vermerkt war, dass es sich um einen Versuch zur automatisierten Datenerhebung im Bereich der Diagnostik handelt.

Als Testinstrument wurden die Untertests WORTAUSWAHL (WA), GEMEINSAMKEITEN (GE) und ZAHLENREIHEN (ZR) des IST 70 (AMTHAUER, 1970) eingesetzt. Dabei wurde die Form A1 für die automatisierte und die Parallelform B2 für die herkömmliche Durchführung verwandt.

[1)] Herrn G. GLAUBITZ sei für die Durchführung der Versuche herzlich gedankt.

Die automatisierte Durchführung wurde über das oben beschriebene System gesteuert. Dabei wurden Anweisungen, Übungsbeispiele und Aufgaben wörtlich - soweit nicht systembedingte Änderungen notwendig waren (z.B. nur Grossbuchstaben, Umlautdarstellung mit "e")- aus der Handanweisung übernommen - vgl. Abb. 5.

DER VERSUCH BESTEHT AUS DREI AUFGABENGRUPPEN. VOR BEGINN JEDER GRUPPE WERDEN DIE DARIN VORKOMMENDEN AUFGABEN BESCHRIEBEN, UND ES WIRD IHNEN AN BEISPIELEN GEZEIGT, WIE SIE GELOEST WERDEN.

SIE KOENNEN WAHRSCHEINLICH IN DER ZUR VERFUEGUNG STEHENDEN ZEIT NICHT ALLE AUFGABEN LOESEN. LASSEN SIE SICH DAVON NICHT BEUNRUHIGEN. BEARBEITEN SIE ALLE AUFGABEN IN DER GEGEBENEN REIHENFOLGE. HALTEN SIE SICH ABER NICHT ZU LANGE BEI EINER ANTWORT AUF.

Abb. 5: Allgemeine Versuchsanweisung für die 1. Versuchsserie.

Die vorgegebenen Zeitbegrenzungen der Untertests entsprachen denen der herkömmlichen Darbietungsform; jedoch wurde hier nicht genau nach Ablauf der Testzeit sondern erst nach abgeschlossener Bearbeitung des entsprechenden Items abgebrochen. Das zuletzt bearbeitete Item wurde nur dann gewertet, wenn bis zum Ende der vorgegebenen Zeit die Taste "Eingabe Ende" betätigt worden war.

Entsprechend den oben aufgeführten Hypothesen ergaben sich für die automatisierte Testdurchführung (Einzelversuch) drei Versuchsbedingungen:

1. Versuchsbedingung (Untertest WA)

Jedes Item wird einzeln und ohne laufende Nummer vorgegeben, der Proband weiss nicht, wieviele Aufgaben er zu bearbeiten hat. Die Möglichkeit zur Eingabekontrolle und zur Eingabekorrektur sind gegeben (Gesamtzeit: 6 Minuten).

2. Versuchsbedingung (Untertest GE)

Diese Versuchsbedingung entspricht 1., nur wird dem Probanden vorher mitgeteilt, dass 20 Aufgaben zu

bearbeiten sind; daneben sind die einzelnen Aufgaben zu seiner Orientierung durchnummeriert (Gesamtzeit: 8 Min.).

3. Versuchsbedingung (Untertest ZR)

Zusätzlich zu der Versuchsbedingung 2. kann der Proband hier die Bearbeitung einzelner Aufgaben zurückstellen, indem er statt einer Antwort die Zahl 9999 eingibt. Die zurückgestellten Items werden ihm nach der Bearbeitung aller Aufgaben in der Standardreihenfolge erneut vorgegeben und können dann unter den selben Bedingungen angegangen werden. Dieser Prozess wird so lange fortgesetzt, bis der Proband alle Items bearbeitet hat, oder bis die Testzeit abgelaufen ist (Gesamtzeit: 10 Min.).

Jeder Proband bearbeitete die drei Untertests in der hier vorgestellten Reihenfolge. Eine Variation in der Abfolge der Versuchsbedingungen war nicht möglich, da dann die Zahl der zu bearbeitenden Aufgaben bekannt gewesen und die Versuchsbedingung 1. nicht mehr anwendbar gewesen wäre.

Die Versuchspersonen wurden nach einer kurzen Einführung im Untersuchungsraum allein gelassen, doch verfolgte der Untersuchungsleiter auf einem weiteren an den Computer angeschlossenen Monitor den Ablauf der Untersuchung.

Zu Beginn des Versuchs (vgl. Abb. 6) wurde der Versuchsperson durch das automatische System eine Kenn-Nr. zur späteren Identifizierung zugewiesen. Dann wurde sie über das Unterprogramm TASUP (vgl. S. 53 - 58) in die Tastatur-Bedienung eingewiesen.

Entsprechend den eben beschriebenen Versuchsbedingungen wurden die Untertests WA, ZR und GE automatisch durchgeführt. Nach Beendigung des Versuchs wurden die Versuchsergebnisse über den Lochkartenstanzer und den Zeilendrucker ausgegeben.

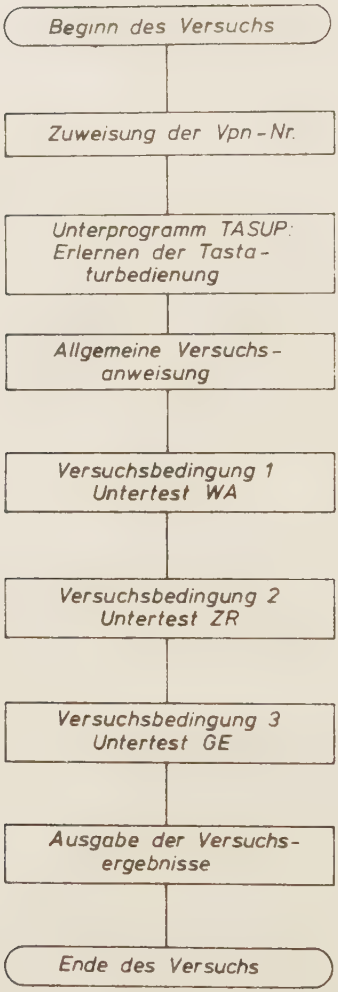


Abb. 6: Schematische Darstellung des Versuchsablaufs in der ersten Versuchsserie.

Die Untersuchung mit der herkömmlichen Testform wurde entsprechend der Handanweisung des IST als Gruppenuntersuchung (Gruppengrösse: 3 - 10 Probanden) durchgeführt. Im Mittel betrug der Abstand zwischen beiden Testsitzungen 7 Tage (Streubreite 2 - 15 Tage). Dabei musste die automatisierte Durchführung vor der herkömmlichen erfolgen, weil sonst wiederum die Versuchsbedingung 1. der automatisierten Form nicht realisierbar gewesen wäre.

Nach beiden Testsitzungen war ein von GLAUBITZ (1975) entwickelter Katalog mit acht ja/nein-Fragen zu beantworten. Die einzelnen Statements, die aufgrund von Befragungen und Verhaltensbeobachtungen bei Vorversuchen zur automatisierten Testdurchführung entstanden sind, lauten:

1) Ich empfand die Versuchssituation als unpersönlich.	NEIN
2) Ich fühlte mich angesprochen.	JA
3) Die Versuchssituation war interessant.	JA
4) Die Versuchssituation war steril.	NEIN
5) Ich fühlte mich gehetzt.	NEIN
6) Die Versuchssituation war angenehm.	JA
7) Die Versuchssituation war kühl/trocken.	NEIN
8) Ich konnte mich gut konzentrieren.	JA

Die hinter den Statements aufgeführte Antwortmöglichkeit (JA bzw. NEIN) gibt die Antwort an, die bei der Berechnung des Gesamtscores für eine positive Beschreibung der Versuchssituation mit einem Punkt bewertet wird.

2. VERSUCHSERGEBNISSE

Die erste Untersuchungsfrage, ob es dem Probanden gelingt, sich auf das System zur automatisierten Testdurchführung einzustellen, kann aufgrund der Beobachtungen des Dialogs zwischen Versuchsperson und Computer zumindest für die hier untersuchte

Stichprobe (Studenten) eindeutig bejaht werden. Der Versuch wurde von keinem Probanden vorzeitig abgebrochen, noch ergaben sich durchführungsbedingte Stockungen während der eigentlichen Untersuchungsphase.

Die Erfahrungen aus ähnlichen Untersuchungen - Z.B. JOHNSON & MIHAL (1973), deren Stichprobe aus Schülern bestand, oder ELWOOD (1972 c), der Versuchspersonen mit niedriger Intelligenz untersuchte - sprechen aber dafür, dass sich auch andere Probandengruppen auf dieses System einstellen können.

A. Übereinstimmung zwischen Testergebnissen aus herkömmlicher und automatisierter Testdarbietung

Um die Frage abzuklären, wieweit die automatisierte Testdurchführung an die herkömmliche anzunähern ist, wurden Produkt-Moment-Korrelationen zwischen den Testwerten in den entsprechenden Untertests bei beiden Durchführungsarten berechnet:

(1) "Überblick" (-) / "Zurückstellen" (-) (WA) r = .378 **[1)]
(2) "Überblick" (+) / "Zurückstellen" (-) (GE) r = .690 ***
(3) "Überblick" (+) / "Zurückstellen" (+) (ZR) r = .719 ***

Bei den weiteren Überlegungen gehen wir davon aus, dass die Versuchspersonen nach der programmierten Einführung in die Tastaturbedingung mit dem automatisierten System vertraut sind. Damit hat die Durchführung des ersten

---

1) Die Kennzeichnung signifikanter Werte erfolgt hier und im folgenden durch "*". Dabei bedeuten:
* : p = 5%; ** : p = 1%; *** : p = 0.1%.

Untertests nicht die Funktion eines "Schock-Absorbers" zur Eingewöhnung in die neuartige Situation, sondern die Ergebnisse des ersten Untertests sind von der Gewöhnung her als vergleichbar mit den aus den beiden anderen Untertests anzusehen.

Aufgrund der Korrelationen können wir feststellen, dass die Variable "Überblick über die bisher geleistete und die noch zu leistende Arbeit" wesentlich zur Annäherung der Ergebnisse aus beiden Untersuchungsformen beiträgt, während durch die zusätzliche Möglichkeit, einzelne Items bei der Bearbeitung zunächst zu überspringen, die Übereinstimmung nur geringfügig höher wird.

Vergleicht man diese Korrelationen zwischen beiden Durchführungsarten mit den von ELWOOD (1972a, 1973), der Koeffizienten von $r > .9$ berichtet, so fallen sie für alle drei Versuchsbedingungen, vor allem aber für die erste, niedriger aus. Dagegen entsprechen die Werte für die zweite und dritte Versuchsbedingung etwa dem von HEDL, O'NEIL & HANSEN (1973), die die Übereinstimmung mit $r = .75$ angeben. Für diese unterschiedlichen Ergebnisse bietet sich eine inhaltliche Erklärung an: Die von ELWOOD vorgenommene Übertragung des WAIS in eine automatisierte Form geschah zum grössten Teil auf mechanischem Wege, während bei HEDL, O'Neil& HANSEN und in diesem Versuch die Testdurchführung computer-gesteuert über ein Datensichtgerät erfolgte. Es lässt sich daher vermuten, dass die grössere äussere Übereinstimmung zwischen beiden Darbietungsformen wesentlich zu den höheren Korrelationskoeffizienten bei ELWOOD beigetragen hat.

Zur weiteren Überprüfung der Unterschiede zwischen den verschiedenen Versuchsbedingungen bei der automatisierten Darbietung und der herkömmlichen Testdurch-

führung gehen wir von den Mittelwerten der Untertests aus; dazu wurden für beide Durchführungsarten die Rohwerte nach der Handanweisung des IST in altersentsprechende Standardwerte transformiert.

Wie Tab. 1 zeigt, sind die t-Tests (abhängige Stichproben) für die Mittelwertsunterschiede zwischen automatisierter und herkömmlicher Darbietung hoch signifikant. Es zeigt sich also für alle drei Versuchsbedingungen eine weit überzufällige Leistungssteigerung in der zweiten Testdurchführung. Dabei werden jedoch die Differenzen zwischen erster und zweiter Durchführung von Versuchsbedingung 1. bis Versuchsbedingung 3. immer geringer.

	Versuchsbed.1		Versuchsbed.2		Versuchsbed.3	
	autom.	herk.	autom.	herk.	autom.	herk.
Mittelwert	100.52	109.86	102.50	109.70	107.60	111.74
Streuung	8.76	8.54	9.87	8.13	11.48	9.71
Diff. MW	9.34		7.20		4.14	
t-Wert	6.79***		7.09***		3.51***	

Tab. 1: Mittelwertsvergleich für abhängige Stichproben zwischen automatisierter und herkömmlicher Testdurchführung in den drei Versuchsbedingungen.

Um nun zu überprüfen, ob sich zwischen den verschiedenen Versuchsbedingungen Unterschiede in der Leistungszunahme von der ersten zur zweiten Testdurchführung zeigen, berechnen wir zunächst für jeden Probanden bei jeder Versuchsbedingung die Differenz zwischen erster und zweiter Durchführung. Über diese Kennwerte für die Leistungszunahme werden nun Mittelwertsvergleiche (t-Tests für abhängige Stichproben) zwischen den drei Versuchsbedingungen berechnet.

	Versuchsbed.1	Versuchsbed.2	Versuchsbed.3
Versuchsbed.1	MW: 8.80 ST: 9.35		
Versuchsbed.2	1.15	MW: 7.00 ST: 6.97	
Versuchsbed.3	2.52*	1.79	MW: 4.34 ST: 8.13

Tab. 2: Mittelwertsvergleich für abhängige Stichproben der Leistungszunahmen zwischen automatisierter und herkömmlicher Testdurchführung in den drei Versuchsbedingungen - in der Diagonalen Mittelwerte und Streuungen der Leistungszunahmen.

Vergleicht man die Leistungszunahmen für die drei Versuchsbedingungen (Untertests) miteinander, so ergibt sich aus Tab. 2, dass für die Zunahmen unter der ersten und dritten Versuchsbedingung ein signifikanter Unterschied besteht, der Zuwachs unter der dritten Bedingung also bedeutsam niedriger ist als unter der ersten Bedingung. Sonst zeigen sich keine überzufälligen Unterschiede.

Wir gehen nun von der Annahme aus, dass bei zweimaliger Testvorgabe von parallelen Formen des IST der übungsbedingte Zuwachs für die Untertests WA, GE und ZR etwa gleich gross ist. Dabei besteht die Möglichkeit, dass die Versuchspersonen bei der automatisierten Durchführung des Untertests WA durchschnittlich etwas schlechtere Ergebnisse erzielen, da es sich um den insgesamt zuerst dargebotenen Test handelt. Doch lässt sich diese Vermutung mit dieser Versuchsanordnung, die - wie oben ausgeführt - eine ausbalancierte Testdarbietung nicht gestattete, nicht überprüfen.

Unter der Voraussetzung eines gleichmässigen Übungsgewinns können wir folgern, dass die unterschiedlichen Leistungszunahmen bei der ersten und dritten Versuchsbedingung auf Unterschiede zwischen diesen beiden Versuchsbedingungen

zurückzuführen sind. Das bedeutet, dass die automatisierte Testdurchführung bei der dritten Versuchsbedingung der herkömmlichen Testdarbietung wesentlich näher kommt als die automatisierte Testdarbietung unter der ersten Versuchsbedingung.

Aufgrund der Ergebnisse bei der Überprüfung der Leistungszunahme können wir hier festhalten: Wenn die Möglichkeiten zur Leistungskontrolle und zum Überspringen einzelner Items gegeben sind, werden die Ergebnisse aus der automatisierten Darbietung denen der herkömmlichen wesentlich ähnlicher, als wenn beide Möglichkeiten nicht oder nur die eine zur Leistungskontrolle gegeben sind.

Nach der Überprüfung der Korrelationen und der Leistungszunahme zwischen beiden Durchführungsarten ergibt sich für die Untersuchungshypothese, dass mit zunehmender Annäherung der automatisierten Darbietung an die herkömmliche die Übereinstimmung zwischen den Ergebnissen immer grösser wird: Es kann als gesichert angenommen werden, dass ohne die beiden Möglichkeiten zur Leistungskontrolle und zum Zurückstellen einzelner Items die Ergebnisse aus beiden Durchführungsformen nicht vergleichbar sind. Ob nur einer oder beide Faktoren für eine hinreichende Übereinstimmung notwendig sind, kann aufgrund der hier vorliegenden Ergebnisse nicht entschieden werden.

B. Auswirkungen der beiden Testsituationen

Die Auswirkungen der beiden unterschiedlichen Darbietungsformen auf die Probanden lassen sich zunächst durch die Summenwerte der positiv von den Probanden beantworteten Statements für die beiden Durchführungsarten überprüfen. In der automatisierten Form liegt der Median der positiven Gesamtscores bei 5, in der herkömmlichen Darbietungsform ist es der Wert 4. Damit ergibt sich, dass beide Testformen

von den Probanden als etwa gleich belastend bzw. nicht-belastend erlebt worden sind.

Etwas genauere Hinweise ergibt die in Tab. 3 vorgenommene Aufschlüsselung der Antworten auf die einzelnen Fragen. Zum schnelleren Überblick über die Statements sind in dieser Tabelle die Inhalte der Fragen noch einmal in verkürzter Form aufgeführt.

Statement	automat. JA	automat. NEIN	herkömml. JA	herkömml. NEIN	in beiden JA	in beiden NEIN	Chi-Qua.
1. unpersönlich	30	20	19	31	17	8	4.0*
2. angesprochen	35	15	27	23	17	5	2.1
3. interessant	43	7	27	23	17	6	21.3***
4. steril	24	26	18	32	11	19	1.1
5. gehetzt	18	32	24	26	9	12	1.0
6. angenehm	29	21	18	32	9	12	4.0*
7. kühl/trocken	37	13	27	23	22	8	3.5
8. konzentrieren	35	15	30	20	21	6	0.7

Tab. 3: Verteilung der Antworten im Fragebogen zur Einschätzung der Versuchssituation und Überprüfung der Gleichverteilungsannahme für beide Versuchsdurchführungen.

Die Überprüfung der Gleichverteilungsannahme der Antworten für die beiden Versuchsdurchführungen erfolgte jeweils mit dem Chi-Quadrat-Test für unkorrelierte Stichproben, da für kein Statement die Vierfelderkorrelation zwischen den Durchführungen signifikante Werte ergab.

Bei der Analyse der Antworten fällt zunächst das Ergebnis für die 3. Frage auf; hier wird die automatisierte Versuchssituation weit überzufällig häufiger als interessant bezeichnet als die herkömmliche. Weiter gilt für die automa-

tisierte Situation, dass sie überzufällig mehr als angenehm (Frage 6), jedoch auch als unpersönlich (Frage 1) von den Probanden erlebt wird.

Von einer höheren psychischen Belastung der Versuchspersonen in einer der beiden Situationen können wir also nicht sprechen. Es lässt sich sogar mit einigen Vorbehalten annehmen, dass die automatisierte Testdarbietung etwas motivierender wirkt, da sie für die Probanden wesentlich interessanter ist. - Doch ist einschränkend zu beachten, dass die hier untersuchte Stichprobe aus Studenten bestand.

Zur Überprüfung der Hypothese, dass die Probanden in der von ihnen bevorzugten Versuchssituation die relativ jeweils besseren Leistungen zeigen, unterteilen wir die Untersuchungsstichprobe in zwei unabhängige Gruppen. Ausgehend von den Werten für die positiven Gesamtbeurteilungen der beiden Versuchssituationen werden die Versuchspersonen entweder der Untergruppe zugeteilt, die die herkömmliche Darbietungsform häufiger positiv beurteilt hat (N = 22), oder derjenigen, die die automatisierte Durchführung häufiger positiv eingeschätzt hat (N = 28).

Um Aussagen über Unterschiede im relativen Leistungszuwachs zwischen diesen beiden Versuchspersonengruppen machen zu können, gehen wir wiederum von den Differenz-Werten für die Leistungszunahme zwischen automatisierter und herkömmlicher Testdarbietung aus. Dabei wird angenommen, dass der Übungsgewinn, der sich bei zweimaliger Testdarbietung aus dem Vertraut-Werden mit den Inhalten der drei Untertests ergibt, für die beiden Untergruppen der Stichprobe (automatisierte Situation bzw. herkömmliche Situation häufiger als positiv beurteilt) im Durchschnitt gleich gross ist. Etwaige Unterschiede in der Leistungszunahme zwischen diesen beiden Untergruppen lassen sich dann nicht auf einen inhaltlich bedingten Übungsgewinn zurückführen, sondern können als Ausdruck der unterschiedlichen Stimulation durch die Versuchssituationen aufgefasst werden.

Die Unterschiede zwischen beiden Versuchspersonengruppen überprüfen wir mit t-Tests für unabhängige Stichproben - vgl. Tab. 4.

		WA	GE	ZR
automatisiert häufiger positiv beurteilt	MW	8.00	5.42	3.75
	ST	10.36	5.98	9.81
herkömmlich häufiger positiv beurteilt	MW	9.81	9.00	5.09
	ST	8.01	7.75	5.44
t-Wert für den MW-Unterschied		.67	1.83	.57

Tab. 4: Mittelwertsunterschiede für den Leistungszuwachs in den drei IST-Untertests zwischen den beiden Versuchspersonengruppen (Einteilungskriterium: häufiger als positiv eingeschätzte Versuchssituation).

Für die drei Untertests liegen die Mittelwerte im Leistungszuwachs zwischen automatisierter und herkömmlicher Testsituation bei denjenigen Probanden niedriger, die die automatisierte Situation häufiger als positiv eingeschätzt haben. Diese Probanden zeigen also im Vergleich zu denen, die die herkömmliche Situation häufiger als positiv beurteilt haben zwischen automatisierter (erster) Testdurchführung und der herkömmlichen (zweiten) Testdarbietung einen geringeren Übungsgewinn. Daraus lässt sich schliessen, dass diese Probanden in der automatisierten Situation stärker stimuliert waren und etwas bessere Leistungen erbrachten. Doch erweist sich keiner dieser Werte als im statistischen Sinne interpretierbar. Unsere Hypothese, dass die Probanden in der von ihnen bevorzugten Versuchssituation bessere Leistungen erbringen, ist damit also nicht bestätigt worden.

3. ZUSAMMENFASSUNG UND FOLGERUNGEN

Während die bisher veröffentlichten Untersuchungen zur automatisierten Testdurchführung implizit davon ausgehen, dass die dort gewählte Form der Übertragung eines herkömmlichen Testverfahrens in die automatisierte Fassung die jeweils angemessene ist, sind wir hier von der inhaltlichen Analyse der Situation bei der herkömmlichen Durchführung eines "Papier-und-Bleistift-Testes" ausgegangen.

Berücksichtigt man die aufgrund der technischen Voraussetzungen notwendigen Beschränkungen bei der vollständigen Automatisierung eines Testverfahrens, so lassen sich für die herkömmliche Testdarbietung zwei wesentliche Variablen aufzeigen, die für die Vergleichbarkeit beider Untersuchungsformen wesentlich sein können:

- Die ständige Kontrollmöglichkeit der schon geleisteten und noch zu leistenden Arbeit.
- Die Möglichkeit, während der Testsituation einzelne Testaufgaben zunächst zu überspringen und sie dann zu einem späteren Zeitpunkt zu bearbeiten.

Aufgrund der experimentellen Überprüfung dieser beiden Variablen können wir feststellen, dass ohne ihre Berücksichtigung die Übereinstimmung der Ergebnisse aus beiden Durchführungsarten nicht zufriedenstellend ist. Ob nur einer von beiden Faktoren oder beide für eine hinreichende Übereinstimmung notwendig sind, konnte nicht genau entschieden werden.

Weiter ergab der Vergleich der Einschätzungen beider Versuchssituationen durch die Probanden, dass keine der beiden als stärker belastend erlebt wurde. Die Vermutung, dass der einzelne Proband, in der von ihm bevorzugten Versuchssituation bessere Ergebnisse erzielt, konnte statistisch nicht bestätigt werden.

Nach der Entwicklung des automatisierten Systems zur Testdurchführung ist damit ein zweites Ziel der vorliegenden Arbeit erreicht: Die mit dem automatisierten System erzielten Testergebnisse sind mit denen aus einer herkömmlichen Testdurchführung hinreichend vergleichbar.

Nun soll in einer zweiten Versuchsserie experimentell überprüft werden, ob aufgrund einiger, bei der automatisierten Testdurchführung zusätzlich anfallender, Informationen diese Form der Testdurchführung der herkömmlichen Testdarbietung in ihrem diagnostischen Wert überlegen ist.

IV. DIE DIAGNOSTISCHE BEDEUTUNG BEI DER AUTOMATISIERTEN TESTDURCHFÜHRUNG ANFALLENDER INFORMATIONEN (2. VERSUCHSSERIE)

Bei der Begründung unseres übergreifenden Untersuchungsziels sind wir davon ausgegangen, dass die Möglichkeiten des Computereinsatzes bei der Testdurchführung bei weitem nicht ausgeschöpft werden, wenn die sich aus der automatisierten Durchführung ergebenden Werte nur im Sinne der in der herkömmlichen Testdurchführung üblichen Richtig/Falsch-Auswertung weiter verwandt werden. Zum Nachweis unserer Hypothese, dass die computer-unterstützte Testdurchführung leicht zusätzliche diagnostische Informationen liefern kann, haben wir in den beiden vorangegangenen Schritten

- zunächst die Entwicklung eines Programmsystems zur automatisierten Durchführung von Tests in "multiple-choice" - Form beschrieben, das die vorhandenen technischen Möglichkeiten berücksichtigt;

- danach (1. Versuchsserie) haben wir aufgezeigt, dass mit diesem System die automatisierte Darbietung eines Verfahrens zur Leistungsmessung ermöglicht wird, und dass die Testergebnisse aus der automatisierten Durchführung mit denen aus der herkömmlichen hinreichend vergleichbar sind, wenn den Probanden zumindest der Überblick über die bisher geleistete und die noch zu leistende Arbeit ermöglicht wird.

Aufbauend auf diese beiden Ergebnisse ist nun die zentrale Untersuchungshypothese zu überprüfen.

1. HERLEITUNG DER UNTERSUCHUNGSFRAGE

Das mit der 2. Versuchsserie verfolgte Ziel lässt sich zusammenfassend etwa kennzeichnen:

Wieweit kann die Messgenauigkeit eines herkömmlichen Testverfahrens zur Leistungsmessung dadurch verbessert werden, dass dieses Instrument in automatisierter Form dargeboten und einige der dabei anfallenden Informationen berücksichtigt werden?

Als mögliche zusätzliche Informationsquellen sind oben zunächst nur global begründet die Latenzzeit bei der Beantwortung einer Testaufgabe und die subjektive Sicherheit des Probanden über die Richtigkeit seiner Antwort eingeführt worden. Bevor wir diese beiden Variablen genauer erläutern, ist noch einmal darauf hinzuweisen, dass

- wir aus einer Reihe möglicher zusätzlicher Informationsquellen diese beiden beispielhaft ausgewählt haben;

- die Messung dieser beiden Variablen prinzipiell nicht nur bei automatisierter Testdurchführung möglich ist, aber durch diese wesentlich erleichtert wird:

  Während die genaue Messung von Latenzzeiten bei der herkömmlichen Testdurchführung mit so erheblichen Schwierigkeiten (z.B. Messfehler bei der Zeitnahme durch den Untersuchungsleiter; übermässige Komplikation der Bedingungen für die Testdurchführung) verbunden ist, dass sie fast schon als Kennzeichen einer automatisierten Darbietung anzusehen ist, lässt sich die Antwortsicherheit auch bei der herkömmlichen Durchführung einschätzen. Für ihre Messung in der automatisierten Form spricht, dass bei dieser Darbietungsart das Verhalten des Probanden genau kontrolliert werden kann; so kann er z.B. nicht die Antwort auf eine Frage "verweigern". (Vgl. den Abschnitt "Software")

A. Die zusätzlichen Informationsquellen

a. Latenzzeit

In einer Reihe von Testverfahren zur Fähigkeitsmessung, die im Einzelversuch durchgeführt werden, wird die vom Probanden zur Bearbeitung einer einzelnen Testaufgabe benötigte Zeit als zusätzliche diagnostische Information herangezogen. Als Beispiele kann man die Untertests "Mosaiktest" oder "Figurenlegen" bei den WECHSLER-Tests nennen. Hier wird die Leistung der Probanden um so höher durch Punkte bewertet, je schneller die Aufgabe gelöst wird. Dabei geht implizit die Annahme ein - explizit wird dieses Punktsystem nicht begründet - dass die Lösungsgeschwindigkeit ein Indikator für Leistungsfähigkeit ist.

Bei computer-unterstützter Testdurchführung bietet es sich geradezu an, zusätzlich zur Antwortregistrierung auch die Zeit zu messen, die der Proband für die Bearbeitung einer Aufgabe benötigt. So ist es nur selbstverständlich, dass in Analogie zu der vielfältigen Erforschung der Latenzzeit, wie sie z.B. in der Wahrnehmungspsychologie erfolgt, hohe Erwartungen in den Aussagewert genau gemessener Zeiten bei der automatisierten Testdurchführung gesetzt werden - z.B. GREEN (1970). Im Gegensatz zu diesen Erwartungen stehen die wenigen bisher bekannt gewordenen Ergebnisse.

Im Bereich der Leistungsmessung zeigen JOHNSON & MIHAL (1973) einen Zusammenhang zwischen Latenzzeit und Gesamttestergebnis auf: je schneller ein Proband das einzelne Item beantwortet, desto höher ist im allgemeinen sein Gesamttestergebnis.

Denkbar ist auch ein "U-förmiger" Zusammenhang zwischen Fähigkeit und Aufgabenschwierigkeit: für die Versuchsperson leichte Aufgaben werden schnell richtig gelöst, Aufgaben von mittlerem Schwierigkeitsgrad benötigen die längste Lösungszeit, und zu schwere Aufgaben werden schnell geraten,

da die Versuchsperson merkt, dass sie auch nach längerem Überlegen nicht auf die Lösung kommen wird. Doch liegen darüber noch keine empirischen Befunde vor.

TOBIAS, HEDL & TOWLE (1974) untersuchten anhand eines Mathematiktests den Zusammenhang zwischen Latenzzeit und Testangst. Es ergab sich entgegen der Untersuchungshypothese, dass ängstliche Probanden bei der Beantwortung schwieriger Items keine längeren Zeiten benötigten als nicht ängstliche Probanden.

Gegen den der Latenzzeit zugeschriebenen diagnostischen Gewinn spricht auch die Untersuchung von DUNN, LUSHENNE & O'NEIL (1972). Bei der automatisierten Durchführung des MMPI ergab sich, dass die zur Beantwortung eines Items benötigte Zeit zu über 50% durch die erforderliche Lesezeit (Länge des Items) erklärt werden kann, während entgegen der Erwartung die Faktoren "Vieldeutigkeit" und "soziale Erwünschtheit" der Testaufgaben nur weitere 6% der Varianz aufklären konnten.

Darüber hinaus liegen zum Problem der Latenzzeitmessung auf dem Gebiet der computer-gesteuerten Testdurchführung keine weiteren Ergebnisse vor. Daher ist kurz noch der Bereich der CAI anzusprechen, wo die Latenzzeitmessung zu den Standardprozeduren gehört. SUPPES & MORNINGSTAR (1970) z.B. benutzen sie zur Abschätzung des Schwierigkeitsgrades zu erlernender Regeln, JUDD & GLASER (1969) als eine Art Lernindex.

b. Antwortsicherheit

Ein Teil des Unbehagens über den beschränkten Informationswert der Antworten auf "Mehrfach-Wahl-Aufgaben" ist darin begründet, dass durch Raten oder Teil-Kenntnisse entstandene Antworten nicht von solchen unterschieden werden können, bei denen der Proband die richtige Lösung genau kennt. Eine Möglichkeit, hier weitere Informationen

vom Probanden zu erhalten, wird dadurch gegeben, ihn aufzufordern, zu jeder Antwort die subjektive Sicherheit für die Richtigkeit dieser Antwort anzugeben.

BOLDT (1974) hebt den dabei zusätzlich auftretenden Gewinn hervor, indem für den Probanden die Testdurchführung dadurch leichter zu akzeptieren wird, dass er seine Unsicherheit mitteilen kann und damit die Angstgefühle geringer werden, die sich aus dem Zwang zur Festlegung auf eine Antwortmöglichkeit ergeben.

Die Nachteile dieser von manchen bevorzugten Vorgehensweise (z.B. SHUFORD, ALBERT & MASSENGILL, 1966) liegen darin, dass mit einem solchen Test nicht nur das Leistungsniveau, sondern z.B. auch die Risikobereitschaft des Probanden erfasst werden. Diese Gefahr lässt sich jedoch zum grössten Teil umgehen, indem nicht nur die Wahrscheinlichkeit für die Richtigkeit einer sondern aller Antwortmöglichkeiten einzuschätzen ist (BOLDT, 1974).

B. Ziel der 2. Versuchsserie

Bezogen auf Testverfahren zur Leistungsmessung in Mehrfach-Wahl-Form erwarten wir also eine Verbesserung der Messgenauigkeit aufgrund der Einbeziehung der beiden Variablen

- o Latenzzeit bei der Beantwortung einer Testaufgabe,
- o subjektive Sicherheit des Probanden über die Richtigkeit der von ihm gegebenen Antwort.

Um zu erfassen, wieweit diese beiden Variablen zur Erhöhung der Messgenauigkeit eines Tests beitragen, wollen wir von folgenden Kennwerten eines Testverfahrens innerhalb der klassischen Testtheorie ausgehen:

- Schwierigkeit der Testaufgaben;
- Trennschärfe der Testaufgaben - definiert als Item-Rohwert-Korrelation (VUKOVICH, 1964);
- Innere Konsistenz des Tests.

Der Versuchsplan für die 2. Versuchsserie umfasst zwei Versuchsgruppen, denen unterschiedliche Möglichkeiten zur Beantwortung der Items vorgegeben werden:

- Versuchsgruppe (1)

  Die zutreffende Antwort ist anzugeben.

- Versuchsgruppe (2)

  Für jede Antwortmöglichkeit ist die subjektive Sicherheit für die Richtigkeit jeder Antwortmöglichkeit anzugeben.

In beiden Versuchsgruppen ist die Latenzzeit bei der Beantwortung jedes einzelnen Items zu messen.

Dabei erwarten wir für die Aufgabenschwierigkeit, dass

- zwischen Beantwortungszeit und Lösungswert bei den einzelnen Items eine SIGNIFIKANTE Korrelation vorliegt;
- die Korrelation in der Versuchsgruppe (1) niedriger ausfallen als in Versuchsgruppe (2);
- über die Gesamtzahl der Items eines Tests eine SIGNIFIKANTE Beziehung zwischen den mittleren Latenzzeiten und den entsprechenden Schwierigkeitsindices besteht.

Für die Trennschärfekoeffizienten wird erwartet, dass

- innerhalb jeder Versuchsgruppe die Korrelation zwischen Item und Rohwert niedriger ist, als wenn in diese Beziehung die Latenzzeiten mit eingehen;

- zwischen beiden Versuchsgruppen die Trennschärfekoeffizienten innerhalb der Versuchsgruppe (2) jeweils höher sind als in Versuchsgruppe (1).

Für die Konsistenz der Tests wird entsprechend erwartet, dass

- innerhalb jeder Versuchsgruppe die Konsistenz mit Einbezug der Latenzzeit höher ist als ohne diese;

- zwischen beiden Versuchsgruppen die Konsistenzkoeffizienten in Versuchsgruppe (2) höher ausfallen als in Versuchsgruppe (1).

C. Versuchsaufbau und Versuchsdurchführung

Der äussere Versuchsaufbau entspricht dem oben (s.S.42-45) beschriebenen Aufbau zur automatisierten Testdurchführung; die Versuchssteuerung erfolgt über das entsprechende System (s.S. 46-58).

Da es in dieser Versuchsserie nicht um die Überprüfung der Übereinstimmung zwischen automatisierter und herkömmlicher Testdurchführung geht, sondern die Messgenauigkeit im Vordergrund steht, ist es notwendig, dass alle Aufgaben von allen Probanden beantwortet werden; d.h. die Untersuchung wird ohne Zeitbegrenzung durchgeführt.

Bei der Bestimmung der Darbietungsform für das einzelne

Item gehen wir von dem oben erbrachten Nachweis (vgl. Abschnitt III) aus, dass das Wissen um die geleistete und noch zu leistende Arbeit eine hinreichende Übereinstimmung zwischen herkömmlicher und automatisierter Darbietung ermöglicht. Wir entschieden uns also in dieser Versuchsserie für die Versuchsbedingung 2. der 1. Versuchsserie, d.h. der Proband weiss wieviele Aufgaben er zu bearbeiten hat, und diese sind zu seiner besseren Orientierung durchnummeriert.

Die Möglichkeit zum Überspringen einzelner Items wurde nicht eingeräumt, weil neben der in der Versuchsgruppe (2) erforderlichen Einschätzung jeder Antwortmöglichkeit nicht noch zusätzliche Verzweigungen gestattet werden sollten, die das Instruktionsverständnis wesentlich erschwert hätten.

Als Versuchsmaterial dienen die drei ersten Untertests des IST 70 (AMTHAUER, 1970), nämlich SATZERGÄNZUNG, WORTAUSWAHL und ANALOGIEN.

Nach der Einführung in die Tastaturbedienung (vgl. Unterprogramm TASUP) wurden die Versuchspersonen durch das Steuerprogramm (vgl. zum weiteren auch Abb. 8) abwechselnd jeweils der ersten oder zweiten Versuchsgruppe zugewiesen.

Den Versuchspersonen beider Versuchsgruppen wurde über den Bildschirm die folgende Versuchsanweisung (vgl. Abb. 7) ausgegeben:

DIE UNTERSUCHUNG BESTEHT AUS 3 AUFGABENGRUPPEN. DIE JEWEILS 20 NACH IHRER SCHWIERIGKEIT GEORDNETE EINZELAUFGABEN ENTHALTEN. VOR BEGINN JEDER AUFGABENGRUPPE WERDEN DIE DARIN VORKOMMENDEN AUFGABEN BESCHRIEBEN, UND ES WIRD IHNEN AN BEISPIELEN GEZEIGT, WIE SIE GELOEST WERDEN.

DIE BEARBEITUNGSZEIT FUER DIE EINZELNEN AUFGABEN UND AUFGABENGRUPPEN IST NICHT BEGRENZT. SIE KOENNEN SICH ALSO IN RUHE MIT JEDER DER AUFGABEN, DIE IHNEN NACHEINANDER EINZELN VORGEGEBEN WERDEN, BESCHAEFTIGEN

BEI ALLEN AUFGABEN HANDELT ES SICH UM 'MEHRFACH-WAHL-AUFGABEN', ZU DENEN JEWEILS FUENF LOESUNGEN VORGEGEBEN SIND. IHRE AUFGABE BESTEHT NATUERLICH DARIN, JEWEILS DIE RICHTIGE LOESUNG ZU FINDEN.

Abb. 7: Allgemeine Versuchsanweisung für die 2. Versuchsserie.

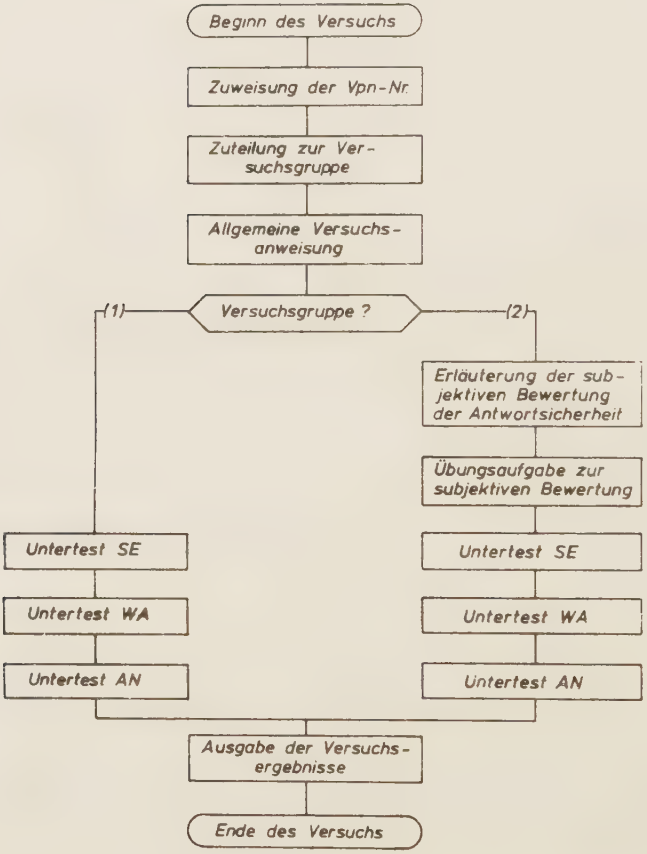


Abb. 8: Schematische Darstellung des Versuchsablaufs in der zweiten Versuchsserie - die zunächst erfolgte Einweisung der Versuchspersonen in die Tastaturbedienung ist hier nicht mehr dargestellt.

Haben die Versuchspersonen angezeigt, dass sie diesen Text (vgl. Abb. 7) gelesen haben, so erhielten die Probanden der Versuchsgruppe (1) die Beschreibung und Beispiele für die erste Aufgabengruppe (Untertest SE) - für beide Versuchsgruppen und jede Aufgabengruppe wurden diese

jeweils wörtlich aus der Handanweisung des IST 70 übernommen - während den Probanden der Versuchsgruppe (2) jetzt die subjektive Bewertung der Antwortrichtigkeit (vgl. Abb. 9) erklärt wurde.

ABER IM GEGENSATZ ZUM SONST UEBLICHEN VORGEHEN, SOLLEN SIE ZU JEDER ANTWORTMOEGLICHKEIT ANGEBEN, OB DIES DIE RICHTIGE LOESUNG SEIN KOENNTE ODER NICHT. DAZU VERTEILEN SIE BITTE BEI JEDER AUFGABE INSGESAMT FUENF PUNKTE SO AUF DIE FUENF LOESUNGSMOEGLICHKEITEN, DASS DIE AUF JEDE ANTWORT ENTFALLENDE PUNKTZAHL DEN GRAD IHRER SICHERHEIT ANGIBT, DASS GERADE DIESE ANTWORT DIE RICHTIGE LOESUNG IST.

WENN SIE ALSO GENAU WISSEN, WELCHE ANTWORT RICHTIG IST, WIRD DIESE LOESUNG FUENF PUNKTE ERHALTEN UND DIE UEBRIGEN VIER BEKOMMEN JEWEILS NULL PUNKTE. - ERSCHEINT IHNEN KEINE DER FUENF LOESUNGEN BESSER ALS DIE VIER ANDEREN, KOENNTEN SIE Z.B. JEDER ANTWORTMOEGLICHKEIT JEWEILS EINEN PUNKT GEBEN.

WIE SIE DIE FUENF PUNKTE AUF DIE FUENF ANTWORTMOEGLICHKEITEN VERTEILEN, STEHT IHNEN NATUERLICH FREI. NUR MUESSEN ZUR BEANTWORTUNG JEDER AUFGABE INSGESAMT FUENF PUNKTE VERGEBEN WERDEN.

Abb. 9: Zusätzliche Versuchsanweisung für die Versuchsgruppe (2) zur Erläuterung der subjektiven Bewertung der Antwortsicherheit.

Danach wurde zusätzlich eine Übungsaufgabe (vgl. Abb.10) aus einem Wortschatztest vorgegeben, die die Probanden nach den hier beschriebenen Antwortregeln zu beantworten hatten. Dabei wurden - wie später auch bei der Testdurchführung - die Punktwerte für die fünf Lösungsmöglichkeiten nacheinander abgefragt.

LOESEN SIE NUN BITTE DIESE DEMONSTRATIONSAUFGABE NACH DEN EBEN BESCHRIEBENEN REGELN.

BERLOCKE
(1)ZWEITERISUR (2) DELIKATESSE (3) SCHMUCKTEIL (4) KOPFBEDECKUNG
(5) GESTEINSART

Abb. 10: Übungsaufgabe zur subjektiven Bewertung der Antwortsicherheiten für die Versuchsgruppe (2).

Die Probanden der Versuchsgruppe (2) erhielten jetzt die Beschreibung und Beispiele für die erste Aufgabengruppe (Untertest SE).

In beiden Versuchsgruppen wurden die Zeiten für die Be-

arbeitung des einzelnen Items und für die Bearbeitung jeder Aufgabengruppe auf ganze Sekunden gerundet gemessen. Die Zeit für die Bearbeitung eines Items (Latenzzeit) ist dabei bestimmt zwischen dem Zeitpunkt der vollständigen Bildschirmausgabe des Aufgabentextes und dem Abschluss der Aufgabenbeantwortung (Betätigung der Taste "Eingabe Ende).

Die Berechnung der Versuchsergebnisse erfolgte aufgrund der jeweils nach Versuchsende auf Lochkarten ausgegebener Daten. Dabei wurden in der Versuchsgruppe (1) die richtigen Lösungen mit einem Punkt bewertet, während in der Gruppe (2) die von dem Probanden der richtigen Lösung zugeteilte Punktzahl als Lösungswert benutzt wurde.

Insgesamt wurden 50 Versuchspersonen (19 männlich, 31 weiblich; vor allem Psychologie-Studenten) untersucht, wobei aufgrund technischer Schwierigkeiten (vor allem Ausfall eines Zeitregisters im Interface) zehn Versuche nicht korrekt durchgeführt werden konnten. Somit besteht jede Versuchsgruppe endgültig aus 20 Probanden.

2. VERSUCHSERGEBNISSE

Bei der Darstellung der Versuchsergebnisse werden unter den Gliederungspunkten

- o Aufgabenschwierigkeit
- o Trennschärfe
- o innere Konsistenz

zunächst innerhalb der beiden Versuchsgruppen die Ergebnisse ohne Latenzzeitmessung denen unter Einbezug der Zeit gegenübergestellt; danach sind jeweils die Ergebnisse aus beiden Versuchsgruppen gegeneinander abzuwägen.

A. Aufgabenschwierigkeit

a. Versuchsgruppe (1)

Zunächst ist zu überprüfen, ob zwischen der richtigen bzw. falschen Antwort auf eine Testaufgabe und der zur Beantwortung einer Testaufgabe benötigten Zeit ein bedeutsamer Zusammenhang besteht.

Item	Satzergänzung	Wortauswahl	Analogien
1	-0.38	-[1)]	-
2	-0.08	0.04	-
3	-0.41	-0.07	-
4	0.04	-0.43*	-0.05
5	0.02	0.19	-0.13
6	0.06	-0.41	-0.03
7	-0.74***	-0.30	0.19
8	0.42	0.08	-0.18
9	0.28	-0.18	-0.57**
10	0.23	-0.06	-0.12
11	-0.10	0.37	-0.07
12	-0.12	0.31	-0.31
13	-0.24	-0.33	0.03
14	-0.43*	0.07	-0.54*
15	0.24	0.33	-0.34
16	-0.09	-0.48*	-0.11
17	0.05	0.14	0.06
18	0.27	0.19	-0.03
19	0.11	-0.05	-0.02
20	-0.18	0.44*	0.04

Tab. 5: Punkt-biseriale Korrelationen zwischen Lösungswert des Items und Latenzzeit bei den drei Untertests des IST in der Versuchsgruppe (1).

1) Falls kein Korrelationskoeffizient angegeben ist, war eine Punkt-biseriale Korrelation nicht zu berechnen, da die entsprechende Aufgabe von allen Probanden gelöst wurde.

Dazu berechnen wir zwischen dem alternativen Merkmal Lösung des Items und der entsprechenden Latenzzeit die Punkt-biseriale Korrelation. Die entsprechenden Koeffizienten mit Angabe des Signifikanzniveaus sind in Tab. 5 dargestellt.

Bei den 60 Testaufgaben der drei Untertests finden wir insgesamt nur 7 Beziehungen zwischen Lösung der Aufgabe und dafür benötigter Zeit, die mindestens auf dem 5%-Niveau als statistisch gesichert gelten können. Wir müssen also für diese Versuchsgruppe unsere Hypothese verwerfen, dass zwischen Latenzzeit und Lösungswert für eine Testaufgabe ein bedeutsamer Zusammenhang besteht.

Gehen wir von den signifikanten Beziehungen zwischen Lösungszeit und Lösungswert aus, so zeigt sich, dass 6 der 7 Koeffizienten negativ sind. Während die Anzahl der negativen und positiven Koeffizienten insgesamt etwa gleich ist, können wir für den Fall, dass ein bedeutsamer Zusammenhang vorliegt, mit einiger Vorsicht erwarten, dass eine kurze Lösungszeit für die richtige Beantwortung der Aufgabe spricht.

Wir wollen nun untersuchen, ob zwischen der Schwierigkeit der Testaufgaben und den durchschnittlich zur Lösung benötigten Latenzzeiten bei den Aufgaben eines Untertests ein bedeutsamer Zusammenhang besteht.

Dazu gehen wir von den Schwierigkeitsindices (P_i) der einzelnen Aufgaben aus, die den prozentualen Anteil der Probanden angeben, die diese Aufgabe richtig gelöst haben, und transformieren P_i in LOGIT P_i (z.B. LORD & NOVICK, 1968), wobei

$$\text{LOGIT } P_i = \ln ((1 - P_i) / P_i) = b_i \ .$$

Damit gewinnen wir zur Aussage über die Schwierigkeit einer Aufgabe einen annähernd normalverteilten Schwierigkeitsparameter b_i.

Zwischen den Schwierigkeitsparametern und den mittleren Latenzzeiten berechnen wir für die Aufgaben eines Untertests die entsprechenden Produkt-Moment-Korrelationen. [1)]

Untertest	r
Satzergänzung	.28
Wortauswahl	.54*
Analogien	.63**

Tab. 6: Produkt-Moment-Korrelationen zwischen den Schwierigkeitsparametern b und den mittleren Latenzzeiten für die Items der drei IST-Untertest in der Versuchsgruppe (1)

Es zeigt sich im ersten Untertest ein im statistischen Sinne zufälliger, in den beiden anderen Untertests jedoch bedeutsamer Zusammenhang zwischen mittlerer Latenzzeit und Schwierigkeitsparametern der Items. Wenn wir davon ausgehen, dass die unterschiedliche Art des Aufgabenmaterials in den drei Untertests keinen Einfluss auf die Beziehung zwischen Schwierigkeit einer Aufgabe und benötigter Lösungszeit ausübt, so können wir hier für die Gesamtheit der Versuchspersonen in Versuchsgruppe (1) ablesen, dass mit zunehmender Gewöhnung an das automatisierte Testinstrument leichtere Items schneller als schwere beantwortet werden. Dieser Zusammenhang lässt sich sehr deutlich in den Abb. 11 bis 13 ablesen.

---

1) Eine Übersicht über die Ergebnisse in der Versuchsgruppe (1) geben die Tab. A1, A2 und A3 im Anhang.

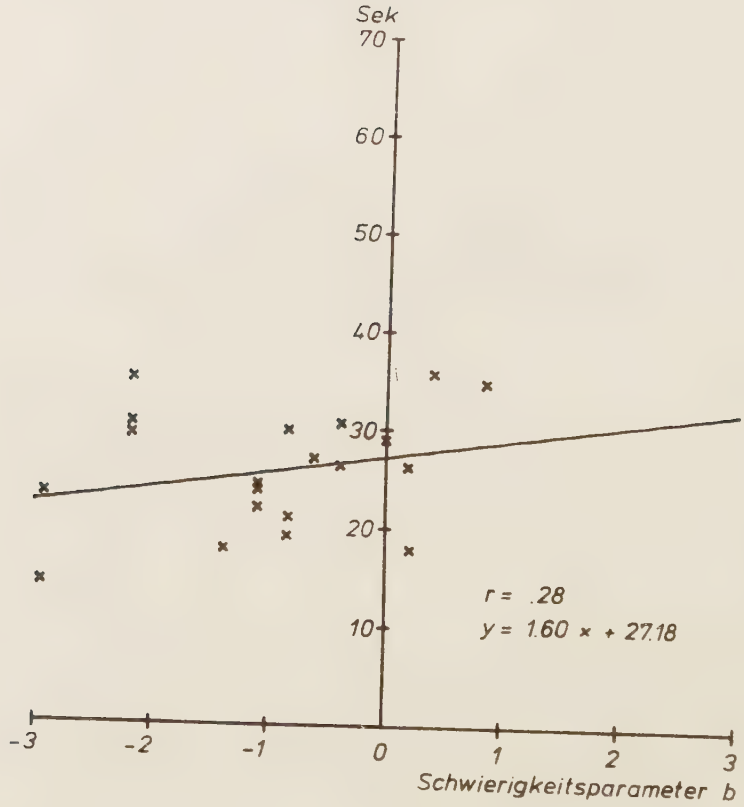


Abb. 11: Beziehung zwischen Aufgabenschwierigkeit und mittlerer Bearbeitungszeit für die Aufgaben des Untertests SATZERGÄNZUNG in der Versuchsgruppe (1)

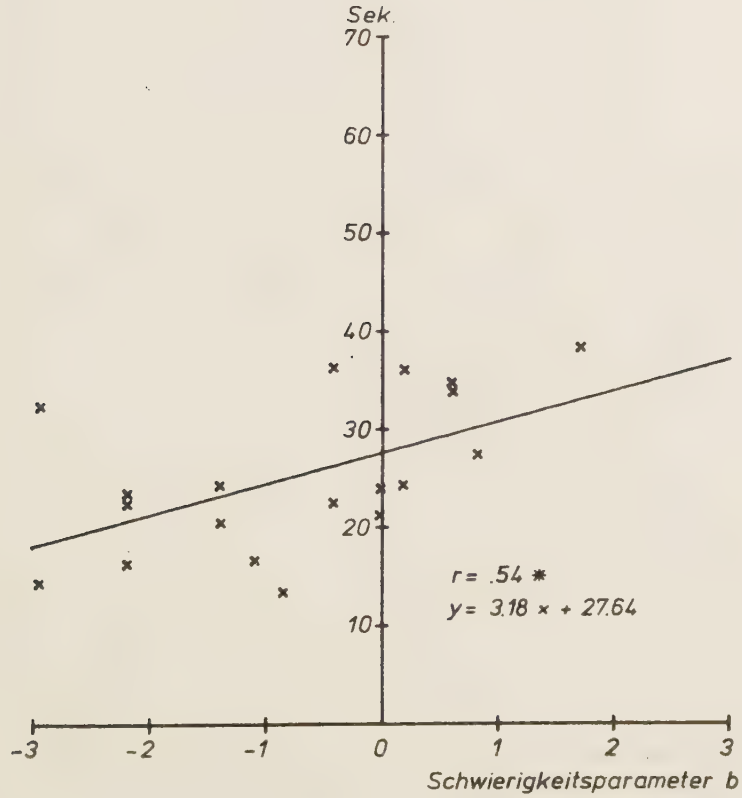


Abb. 12: Beziehung zwischen Aufgabenschwierigkeit und mittlerer Bearbeitungszeit für die Aufgaben des Untertests WORTAUSWAHL in der Versuchsgruppe (1)

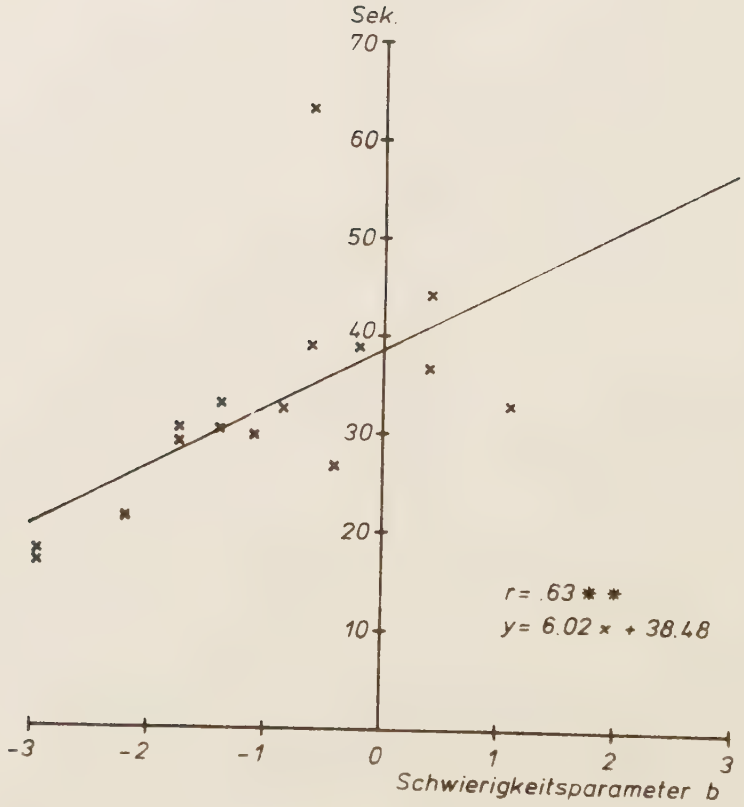


Abb. 13: Beziehung zwischen Aufgabenschwierigkeit und mittlerer Bearbeitungszeit für die Aufgaben des Untertests ANALOGIEN in der Versuchsgruppe (1)

b. Versuchsgruppe 2

Auch in dieser Versuchsgruppe, in der die Probanden zu jeder Antwortmöglichkeit eines Items in Punktform die Wahrscheinlichkeit angeben sollten, dass gerade diese Lösung die richtige sei, überprüfen wir zunächst, ob zwischen dem Antwortwert der Testaufgabe und der zur Beantwortung benötigten Zeit ein bedeutsamer Zusammenhang besteht.

Item	Satzergänzung	Wortauswahl	Analogien
1	0.09	-0.95***	-[1]
2	-0.14	-0.44*	-0.08
3	-0.05	-0.35	-0.19
4	-0.22	-0.03	-0.37
5	-0.31	-0.53*	-0.29
6	-0.23	-0.13	-0.14
7	-0.71***	0.21	-0.13
8	-0.08	-0.23	-.052*
9	-0.14	-0.33	-0.66*
10	-0.18	-0.02	-0.14
11	-0.07	0.04	-0.02
12	-0.47*	-0.11	0.10
13	-0.26	-0.18	-0.02
14	-0.15	0.09	-0.03
15	-0.13	-0.16	-0.33
16	-0.29	-0.53*	-0.42
17	-0.28	-0.33	0.02
18	-0.28	0.21	-0.42
19	-0.16	-0.31	-0.15
20	0.12	0.10	-0.49*

Tab. 7: Produkt-Moment-Korrelationen zwischen dem Punktwert für die Lösung des Items und Latenzzeit bei den drei Untertests des IST in der Versuchsgruppe (2)

---

1) Hier konnte keine Korrelation berechnet werden, da die Punktwerte keine Varianz aufweisen.

Dazu werden für jedes Item zwischen den auf die richtige Lösung entfallenen Punktwerten und der entsprechenden Latenzzeit Produkt-Moment-Korrelationen berechnet.

Bei überwiegend negativen Korrelationen[1] zwischen Lösung und Latenzzeit finden sich bei 9 von 59 Items signifikante bzw. hoch signifikante Beziehungen. Insgesamt ist damit die Hypothese, dass zwischen der Lösung des Items und der dafür benötigten Zeit ein bedeutsamer Zusammenhang besteht, zurückzuweisen. Wir können jedoch festhalten, dass unter der Bedingung eines bedeutsamen Zusammenhanges sehr wahrscheinlich bei schnellerer Bearbeitungszeit die Aufgabe besser gelöst wird.

Zur Überprüfung des Zusammenhanges zwischen Schwierigkeit der Testaufgaben und der durchschnittlich zur Lösung benötigten Zeit berechnen wir zunächst die Schwierigkeitsindices der einzelnen Testaufgaben und transformieren diese wiederum in die Schwierigkeitsparameter b. - Bei dem in der Versuchsgruppe (2) durchgeführten Antwortsystem mit der subjektiven Bewertung der einzelnen Lösungsmöglichkeiten ist der Schwierigkeitsindex definiert als das Verhältnis aus der Summe der auf die richtige Lösung entfallenden Punkte zu der Gesamtsumme der Punkte.[2]

Die Beziehungen zwischen den Schwierigkeitsparametern b und den durchschnittlichen Lösungszeiten sind in Abb.14-16 dargestellt.

---

1) Eine Übersicht über Schwierigkeitsparameter der Items Mittelwerte und Streuungen der Punktwerte und Latenzzeiten geben die Tab. A4, A5 und A6 des Anhangs.

2) In Tab. A7, A8 und A9 des Anhanges sind für die drei Untertests die subjektiven Wahrscheinlichkeiten für die Richtigkeit jeder Lösungsmöglichkeit einer Testaufgabe zusammengestellt.

Die Beziehungen zwischen den Schwierigkeitsparametern b und den durchschnittlichen Lösungszeiten sind in Abb. 14-16 dargestellt. Wir fassen hier noch einmal die Korrelationskoeffizienten für die drei Untertests zusammen.

Untertest	r
Satzergänzung	.17
Wortauswahl	.37
Analogien	.59**

Tab. 7: Produkt-Moment-Korrelationen zwischen den Schwierigkeitsparametern b und den mittleren Latenzzeiten für die Items der drei IST-Untertests in der Versuchsgruppe (2)

Während für die Untertests Satzergänzung und Wortauswahl der Zusammenhang zwischen den Schwierigkeitsparametern der einzelnen Testaufgaben und der durchschnittlichen Lösungszeit statistisch nicht zu sichern ist, finden wir dafür eine sehr signifikante Beziehung beim Untertest Analogien. Wir können feststellen, dass über die drei Untertests die Korrelationskoeffizienten ansteigen. Wenn wir wiederum davon ausgehen, dass der Zusammenhang zwischen Aufgabenschwierigkeit und benötigter Lösungszeit von der inhaltlichen Art der vorgegebenen Testaufgaben unabhängig ist, zeigt sich mit zunehmender Gewöhnung an die automatisierte Testdarbietung und besonders an das ungewohnte Antwortschema für die Versuchsgruppe (2), dass im Durchschnitt eine Testaufgabe um so schneller beantwortet wird, je leichter sie ist. (Vgl. bes. Abb. 16)

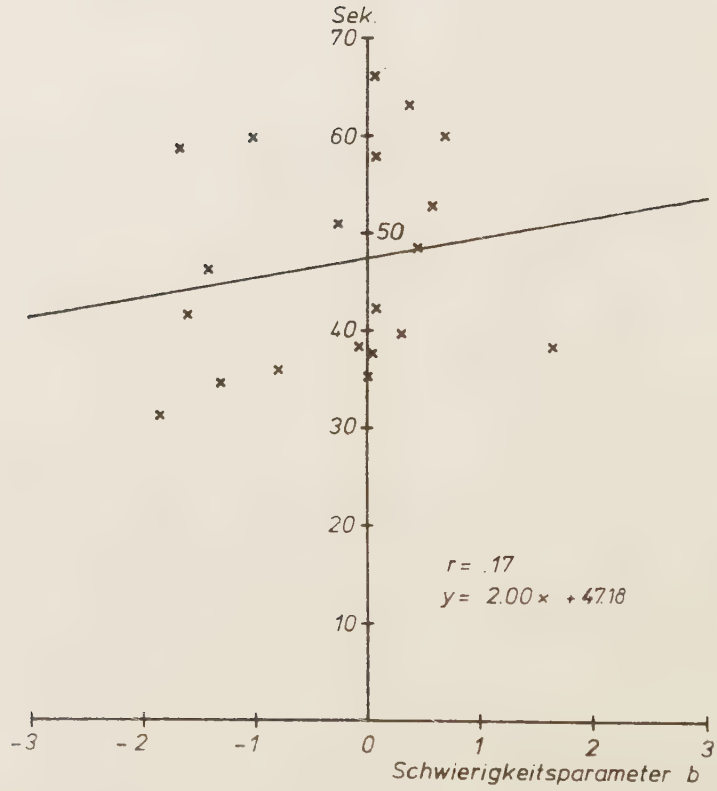


Abb. 14: Beziehung zwischen mittlerem Punktwert und mittlerer Bearbeitungszeit für die Aufgaben des Untertests SATZERGÄNZUNG in der Versuchsgruppe (2)

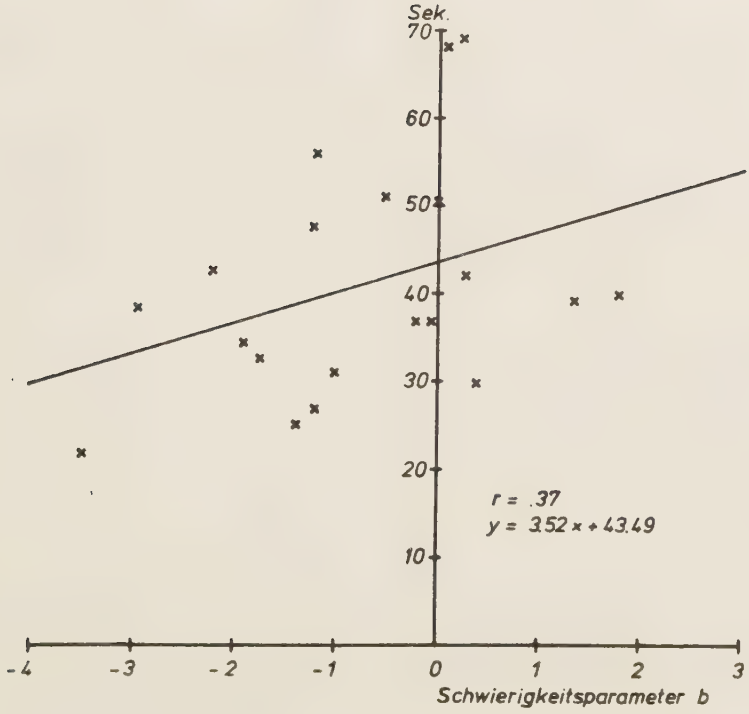


Abb. 15: Beziehung zwischen mittlerem Punktwert und mittlerer Bearbeitungszeit für die Aufgaben des Untertests WORTAUSWAHL in der Versuchsgruppe (2)

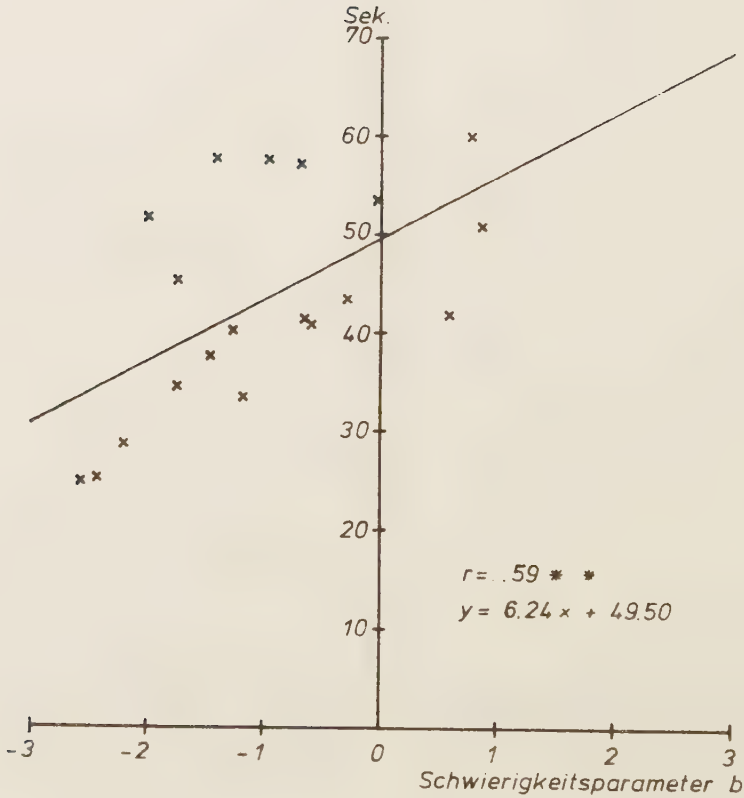


Abb. 16: Beziehung zwischen mittlerem Punktwert und mittlerer Bearbeitungszeit für die Aufgaben des Untertests ANALOGIEN in der Versuchsgruppe (2)

c. Vergleich der beiden Versuchsgruppen

Unter dem Oberbegriff der Aufgabenschwierigkeit haben wir bisher getrennt für jede der beiden Versuchsgruppen, die sich durch die Antwortmöglichkeit auf Items in "multiple-choice"-Form unterscheiden, den Zusammenhang zwischen der Lösung einer Aufgabe und der dafür benötigten Zeit überprüft. Dabei ergab sich für beide Versuchsgruppen, dass insgesamt keine Beziehung zwischen diesen beiden Variablen besteht, es aber für den Fall, dass eine solche Beziehung besteht, sehr wahrscheinlich ist, dass eine kurze Latenzzeit für die Richtigkeit der Aufgabenbeantwortung spricht.

Es ist jetzt zu fragen, ob die Beziehung zwischen Richtigkeit der Antwort und Latenzzeit unter Berücksichtigung der beiden Antwortmodi gleich bleibt. Wir transformieren dazu die Korrelationskoeffizienten r in "normalisierte" z'-Werte, wobei

$$z' = .5 \ln (1 + r) / (1 - r)$$

ist und prüfen die Differenz zwischen den beiden z'-Werten eines Items über den z-Test. Diese Differenzen mit Angabe des entsprechenden Signifikanzniveaus sind in Tab. 8 zusammengestellt.

Insgesamt finden wir nur eine Testaufgabe (WA, Item 5), bei dem sich die Korrelationskoeffizienten zwischen Lösung und Bearbeitungszeit statistisch bedeutsam zwischen den beiden Versuchsgruppen unterscheiden. Bei dieser Aufgabe findem wir in der Versuchsbedingung (1) einen nicht signifikanten posiven Koeffizienten, während der negative unter Bedingung (2) auf dem 5%-Niveau signifikant ist.

Item	Satzergänzung	Wortauswahl	Analogien
1	.49	-[1)]	-
2	.06	.52	-
3	.39	.29	-
4	.26	.43	.33
5	.34	.72*	.15
6	.36	.31	.09
7	.07	.52	.32
8	.53	.31	.39
9	.43	.16	.12
10	.41	.05	.03
11	.03	.43	.05
12	.38	.43	.42
13	.03	.17	.05
14	.31	.02	.58
15	.37	.51	.01
16	.21	.07	.33
17	.33	.48	.04
18	.56	.02	.41
19	.28	.27	.13
20	.31	.37	.50

Tab. 8: Differenz (absoluter Betrag) zwischen den z'-Werten der Korrelationskoeffizienten aus Versuchsgruppe (1) und Versuchsgruppe (2) für den Zusammenhang von Lösung und Bearbeitungszeit der Testaufgaben in den drei IST-Untertests

Vergleichen wir nun die beiden Versuchsgruppen darauf, bei welchen Items in beiden eine bedeutsame Beziehung zwischen Richtigkeit der Antwort und Lösungszeit besteht, so ergeben sich 3 Übereinstimmungen (SE: 7; WA: 16; AN: 9).

---

1) Bei diesen Items war keine Differenzbildung zwischen den Koeffizienten aus beiden Versuchsgruppen möglich, da in einer oder beiden Gruppen die Korrelationen nicht zu berechnen waren.

Entgegen unserer Untersuchungshypothese, dass die Beziehungen zwischen Lösung eines Items und entsprechender Latenzzeit in der Versuchsgruppe (2) deutlicher hervortreten, müssen wir feststellen, dass der Antwortmodus keinen Einfluss auf die Höhe der Beziehung hat.

Für den Zusammenhang zwischen mittlerer Bearbeitungszeit und Schwierigkeit der Items eines Untertests hatte sich unter der Voraussetzung, dass die Art des Aufgabenmaterials keinen Einfluss auf diese Beziehung hat, ergeben, dass in beiden Versuchsgruppen mit zunehmender Gewöhnung an die Art der automatisierten Darbietung eine stärkere Korrelation besteht. Um zu überprüfen, ob sich diese Korrelationen zwischen den beiden Versuchsgruppen unterscheiden, gehen wir wiederum von der Differenz der z'-Werte der Koeffizienten aus.

Untertest	V.-Gruppe (1)	(2)	Differenz
Satzergänzung	.29	.17	.12
Wortauswahl	.60	.39	.21
Analogien	.74	.68	.06

Tab. 9: z'-Werte und Differenzen der z'-Werte zwischen den beiden Versuchsgruppen für die Korrelationen aus Schwierigkeitsparametern und mittlerer Bearbeitungszeit für die Items der drei IST-Untertests

Die jeweiligen Korrelationskoeffizienten für die drei Untertests in den beiden Versuchsgruppen unterscheiden sich statistisch nicht bedeutsam; die Art der Antwortmöglichkeit wirkt sich hier also nicht aus.

Insgesamt können wir für den Komplex der Aufgabenschwierigkeit festhalten, dass

- unabhängig von dem Antwortmodus keine durchgängige Beziehung zwischen der Richtigkeit des Ergebnisses und der Latenzzeit besteht;

- o wenn eine solche Beziehung vorliegt, die Wahrscheinlichkeit grösser ist, bei kürzerer Bearbeitungszeit eine richtige Lösung zu finden;
- o unter der Voraussetzung, dass die Art des Aufgabenmaterials keinen Einfluss auf die Beziehung zwischen mittlerer Bearbeitungszeit und Aufgabenschwierigkeit besteht, mit zunehmender Gewöhnung an die Form der automatisierten Testdarbietung die Korrelation zwischen diesen beiden Variablen höher wird;
- o der Antwortmodus auf die Beziehung zwischen Aufgabenschwierigkeit und mittlerer Bearbeitungszeit keinen Einfluss hat.

B. Trennschärfe

Das wesentlichste Kriterium für die Güte eines einzelnen Testitems ist in der klassischen Testtheorie der Trennschärfekoeffizient; als testinternes Validitätskriterium ermöglicht er eine Aussage darüber, wie genau sich aus der Beantwortung einer Testaufgabe der Gesamttestwert vorhersagen lässt. Rechnerisch handelt es sich um die Korrelation zwischen den Antworten auf eine Testaufgabe und den Gesamtrohwerten, vgl. z.B. LIENERT (1969).

Die bei der automatisierten Testdurchführung anfallende Möglichkeit zur genauen Messung der Latenzzeiten bei der Beantwortung des einzelnen Items soll nun daraufhin untersucht werden, ob unter Benutzung dieser Variablen eine genauere Vorhersage des Gesamttestergebnisses zu erreichen ist. Dazu gehen wir von der multiplen Korrelation der beiden unabhängigen Variablen "Beantwortung des Items"

und "Latenzzeit" mit der abhängigen Variablen "Gesamttestwert" aus und berechnen diese als einen "multiplen Trennschärfekoeffizienten" über:

$$r_{g.al} = \sqrt{\frac{r^2_{ga} + r^2_{gl} - 2r_{ga}r_{gl}r_{al}}{1 - r^2_{al}}}$$

Dabei bedeuten:

$r_{g.al}$: multipler Trennschärfekoeffizient von Antwort und Latenzzeit gegenüber dem Gesamttestwert,

r_{ga} : Korrelation zwischen Gesamttestwert und Antwort,

r_{gl} : Korrelation zwischen Gesamttestwert und Latenzzeit,

r_{al} : Korrelation zwischen Antwort und Latenzzeit.

a. Versuchsgruppe (1)

Beim Untertest "Satzergänzung" ergeben sich folgende multiple Trennschärfekoeffizienten:

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	.08	.51*	-.38	.58**
2	.38	.43	-.08	.60**
3	.43	.18	-.41	.57*
4	.25	.44	.05	.49*
5	-.03	.15	.03	.15
6	.34	.43	.06	.53*
7	-.04	-.01	-.74***	.06
8	.61**	.64**	.43	.73***
9	.34	.14	.28	.34
10	.40	.47*	.23	.55*
11	.11	.09	-.10	.14
12	.10	.18	-.13	.21
13	.20	.42	-.24	.52*
14	-.19	.14	-.43	.19
15	.51*	.10	.24	.51*
16	.31	.34	-.10	.48*
17	.21	.08	.06	.21
18	.48*	.65***	.27	.72***
19	.47*	.45	.12	.61**
20	.18	.51*	-.19	.58**

Tab. 10: Multiple Trennschärfekoeffizienten ($r_{g.al}$) für die Items des Untertests "Satzergänzung" aufgrund der Korrelation Antwort/Gesamtwert (r_{ga}), Latenzzeit/Gesamtwert (r_{gl}) und Antwort/Latenzzeit (r_{al}) in der Versuchsgruppe (1)

Wie ein erster Vergleich zwischen den üblichen Trennschärfekoeffizienten[1] (r_{ga}) und den Korrelationen zwischen Latenzzeiten und Gesamttestergebnis (r_{gl}) zeigt, weisen beide etwa ähnliche Beziehungen zum Gesamttestergebnis auf. Eine Berechnung der durchschnittlichen Korrelation - nach Transformation in z'-Werte - ergibt für $\bar{r}_{ga}$ = .29 und $\bar{r}_{gl}$ = .33.

Gehen wir von der Signifikanz der Korrelationen aus, so sind die multiplen Trennschärfekoeffizienten ($r_{g.al}$) häufiger statistisch bedeutsam, als die herkömmlichen. Für den durchschnittlichen multiplen Trennschärfekoeffizienten ergibt sich $\bar{r}_{g.al}$ = .46. Dieser Wert unterscheidet (z = 2.55**) sich sehr bedeutsam von dem durchschnittlich herkömmlichen Trennschärfekoeffizienten ($\bar{r}_{ga}$ = .29). Damit ergibt sich für den Untertest SE eine bedeutsame Verbesserung der Trennschärfe, wenn als zusätzliche Variable die Latenzzeit für die Beantwortung der einzelnen Items einbezogen wird.

Beim Untertest WA (vgl. Tab. 11) erhalten für die Korrelationen als durchschnittliche Werte $\bar{r}_{ga}$ = .29, $\bar{r}_{g.al}$=.32. Wie die sehr niedrige durchschnittliche Beziehung zwischen Latenzzeiten und Gesamttestwert ($\bar{r}_{gl}$ = .11) erwarten lässt, finden wir nur eine geringfügige (z -Wert zur Prüfung des Unterschieds: .42) Erhöhung der durchschnittlichen multiplen Trennschärfe. Das bestätigt auch ein Vergleich der einzelnen Koeffizienten.

---

1) Ein Vergleich mit dem Trennschärfekoeffizienten der Standardisierungsstichprobe des IST ist nicht möglich, da AMTHAUER (1970) weder die einzelnen noch durchschnittliche Werte angibt.

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	-[1]	-	-	-
2	.21	-.06	.04	.22
3	.13	.08	-.08	.16
4	.14	-.02	-.43	.15
5	.45	.29	.19	.50*
6	-.17	.08	-.42	.17
7	-.12	.07	-.30	.12
8	.22	-.22	.08	.33
9	.30	.09	-.18	.34
10	.23	.05	-.07	.24
11	.38	.17	.38	.38
12	.49*	-.05	.31	.53*
13	.24	-.08	-.34	.24
14	.33	-.07	.08	.34
15	.39	.08	.34	.40
16	.20	.04	-.48	.26
17	.45	-.17	.14	.50*
18	-.14	.12	.20	.20
19	.29	.10	-.06	.31
20	.50*	.28	.44	.50*

Tab. 11: Multiple Trennschärfekoeffizienten für die Items des Untertests "Wortauswahl" aufgrund der Korrelationen Antwort/Gesamtwert, Latenzzeit/Gesamtwert und Antwort/Latenzzeit in der Versuchsgruppe (1). - Notationen s. Tab. 10

---

1) Korrelationsberechnung nicht möglich, da dieses Item von allen Versuchspersonen richtig beantwortet wurde.

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	-[1)]	-	-	-
2	-	-	-	-
3	-	-	-	-
4	.42	.08	-.05	.43
5	.39	.16	-.14	.44
6	-.12	-.37	-.03	.39
7	-.36	-.25	.19	.40
8	.08	.18	-.18	.21
9	.15	.08	-.58**	.26
10	.56*	-.07	-.12	.56*
11	.67***	-.03	-.07	.67***
12	.20	.31	-.32	.44
13	.42	.0	.04	.42
14	.46*	-.24	-.54*	.46*
15	.28	.04	-.35	.31
16	.54*	.07	-.12	.56*
17	-.34	.34	.06	.49*
18	.18	.08	-.03	.20
19	.39	.21	-.02	.44
20	.37	.25	.04	.44

Tab. 12: Multiple Trennschärfekoeffizienten für die Items des Untertests "Analogien" aufgrund der Korrelationen Antwort/Gesamtwert, Latenzzeit/Gesamtwert und Antwort/Latenzzeit in der Versuchsgruppe (1) Notationen s. Tab. 10

Ein Vergleich der Signifikanzen der herkömmlichen und multiplen Trennschärfekoeffizienten ergibt im Untertest AN (vgl. Tab. 12) keinen Hinweis auf eine deutliche Erhöhung bei Einbeziehung der Latenzzeiten. Darauf weisen auch die

---

[1)] Korrelationsberechnung nicht möglich.

durchschnittlichen Korrelationen ($\bar{r}_{ga}$ = .36, $\bar{r}_{gl}$ = .16 und $\bar{r}_{g.al}$ = 43) hin. Hier zeigt sich, dass die durchschnittliche Beziehung zwischen Bearbeitungszeit für das einzelne Item und Gesamttestergebnis so gering ist, dass sie nur wenig zur Erhöhung der multiplen Trennschärfen beitragen kann (z -Wert zur Prüfung des Unterschieds zwischen herkömmlich und multipler Trennschärfe: 1.00).

Für einen Vergleich zwischen herkömmlicher und multipler Trennschärfe in den drei Untertests bei der Versuchsgruppe (1) gehen wir noch einmal von den durchschnittlichen Koeffizienten und den z -Werten zur Unterschiedsprüfung aus:

Untertest	Trennschärfe		z - Wert
	herkömmlich	multipel	
Satzergänzung	.29	.46	2.58**
Wortauswahl	.29	.32	.42
Analogien	.36	.43	1.00

Tab. 13: Durchschnittliche herkömmliche und multiple Trennschärfekoeffizienten sowie z -Werte zur Unterschiedsprüfung für die drei IST-Untertests in der Versuchsgruppe (1)

In der Versuchsgruppe (1), bei der die Versuchspersonen zu jeder Testaufgabe jeweils die richtige Lösung anzugeben hatten, finden sich bei zwei der drei Untertests nur geringfügige Erhöhungen für die durchschnittlichen multiplen Trennschärfen, wenn zur Vorhersage des Gesamttestergebnisses nicht nur die Antworten auf die einzelnen Items sondern auch die dafür benötigten Bearbeitungszeiten herangezogen werden. Unsere Hypothese ist damit nur im Untertest Satzergänzung bestätigt worden.

b. Versuchsgruppe (2)

Wie Tab. 14 zeigt, erhöht sich die Anzahl der signifikanten Trennschärfekoeffizienten unter Einbeziehung der Latenzzeiten zwar nur um eins, doch zeigt sich eine gewisse Überlegenheit der multiplen Koeffizienten. Darauf weisen auch die durchschnittlichen Koeffizienten $\bar{r}_{ga}$ = .36, $\bar{r}_{gl}$ = .28 und $\bar{r}_{g.al}$ = .46 des Untertests "Satzergänzung" hin (z -Wert für den Unterschied zwischen herkömmlichen und multiplen Trennschärfe: 1.56; p = 12%).

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	-.14	-.30	.09	.32
2	.51*	-.25	-.14	.54*
3	.34	-.36	-.05	.48*
4	.31	-.28	-.22	.38
5	.59**	-.48*	-.31	.66***
6	.36	-.34	-.23	.45
7	.29	-.31	-.71***	.33
8	.23	-.36	-.08	.42
9	-.08	-.29	-.14	.32
10	.22	-.43	-.18	.45
11	.57*	.19	-.07	.61**
12	.34	-.26	-.47*	.36
13	-.12	-.21	-.26	.27
14	-.38	.01	-.15	.38
15	.36	.21	-.13	.45
16	.41	-.38	-.29	.49*
17	.49*	-.54*	-.28	.65***
18	.46*	.07	-.28	.50*
19	.43	.07	-.16	.45
20	.46*	.15	.12	.47*

Tab. 14: Multiple Trennschärfekoeffizienten für die Items des Untertests "Satzergänzung" aufgrund der Korrelationen Antwort/Gesamtwert, Latenzzeit/Gesamtwert und Antwort/Latenzzeit in der Versuchsgruppe (2) - Notationen s. Tab. 10

Die entsprechenden Ergebnisse für den Untertest WA finden sich in Tab. 15; ein Vergleich der signifikanten Trennschärfekoeffizienten ohne und mit Einbezug der Variablen "Latenzzeit" ergibt, dass in der multiplen Form zusätzlich die Koeffizienten von zwei Items signifikant werden. Doch handelt es sich dabei um solche Koeffizienten, die schon in der herkömmlichen Form knapp unter dem entsprechenden Signifikanzniveau lagen. Dies zeigen auch die durchschnittlichen Trennschärfekoeffizienten ($\bar{r}_{ga}$ = .28, $\bar{r}_{gl}$ = .20, $\bar{r}_{g.al}$ = .37); die geringe Beziehung zwischen Latenzzeit und Gesamttestergebnis trägt nur wenig zur Erhöhung des durchschnittlichen herkömmlichen im Vergleich zum multiplen Trennschärfekoeffizienten bei (Unterschiedsprüfung: z = 1.30).

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	.48*	-.50*	-.95***	.50*
2	.02	-.33	-.44	.36
3	.15	.11	-.35	.23
4	.35	-.15	-.03	.37
5	.44	-.08	-.53*	.48*
6	.16	-.24	-.13	.27
7	.12	.12	.21	.15
8	.08	-.07	-.23	.10
9	.06	-.22	-.33	.22
10	-.02	-.05	-.02	.05
11	.29	.23	.04	.36
12	.61**	.12	-.11	.64**
13	.34	-.25	-.18	.38
14	.44	.26	.09	.49*
15	.55*	.10	-.16	.58**
16	.57*	-.30	-.53*	.57*
17	.09	-.07	-.33	.10
18	-.17	-.26	.21	.34
19	.23	.35	-.31	.49*
20	.23	.26	.10	.33

Tab. 15: Multiple Trennschärfekoeffizienten für die Items des Untertests "Wortauswahl" aufgrund der Korrelationen Antwort/Gesamtwert, Latenzzeit/Gesamtwert und Antwort/Latenzzeit in der Versuchsgruppe (2) - Notationen s. Tab. 10

Die Trennschärfebestimmung mit der zusätzlichen Berücksichtigung der Variablen "Latenzzeit" für die Items des Untertests AN (vgl. Tab. 16) erhöht die Zahl der signifikanten Korrelationen gegenüber den Trennschärfekoeffizienten im herkömmlichen Sinn um zwei; dabei stützt sich die Trennschärfe einer Testaufgabe (Item Nr. 3) nur auf die Beziehung zwischen Latenzzeit und Gesamttestergebnis.

Item	r_{ga}	r_{gl}	r_{al}	$r_{g.al}$
1	-[1]	-	-	-
2	.54*	-.29	-.08	.60**
3	.09	-.53*	-.19	.53*
4	.32	-.50*	-.37	.52*
5	.64**	-.20	-.29	.64**
6	.55*	.0	-.14	.55*
7	.52*	-.32	-.13	.58*
8	.25	-.02	-.52*	.28
9	.46*	-.25	-.66***	.46*
10	.27	-.14	-.14	.28
11	.34	.04	-.02	.34
12	.52*	.03	.10	.52*
13	.54*	-.15	-.02	.55*
14	.32	.15	-.03	.35
15	-.01	.26	-.33	.27
16	.41	-.12	-.42	.41
17	.23	.27	.02	.35
18	.25	.06	-.42	.31
19	.26	.25	-.15	.39
20	.43	-.28	-.49*	.44

Tab. 16: Multiple Trennschärfekoeffizienten für die Items des Untertests "Analogien" aufgrund der Korrelationen Antwort/Gesamtwert, Latenzzeit/Gesamtwert und Antwort/Latenzzeit in der Versuchsgruppe (2) - Notationen s. Tab. 10

1) Eine Korrelationsberechnung war nicht möglich, da dieses Item von allen Probanden richtig gelöst wurde.

Insgesamt finden wir jedoch nur eine geringfügige Erhöhung der durchschnittlichen multiplen Trennschärfe; der mittlere Wert für die herkömmliche Trennschärfe ($\bar{r}_{ga}$ = .38) wird unter Einbeziehung der mittleren Korrelation zwischen Latenzzeit und Gesamtergebnis ($\bar{r}_{gl}$ =.21) nur unwesentlich zum multiplen Trennschärfekoeffizienten ($\bar{r}_{g.al}$ = .45) erhöht (z -Wert: 1.08).

Überblicken wir die Ergebnisse in der zweiten Versuchsgruppe, in der die Probanden die Wahrscheinlichkeit der Richtigkeit ihrer Antworten mit einzuschätzen hatten, finden wir ausgehend vom Vergleich der Trennschärfebestimmung über die herkömmliche Beziehung Antwort/Gesamtergebnis und der multiplen Koeffizienten unter Einbeziehung der zur Aufgabenbeantwortung benötigten Zeit nur eine unwesentliche Vermehrung der statistisch bedeutsamen Koeffizienten. Noch deutlicher wird dieser Sachverhalt, wenn wir von den in Tab. 17 zusammengestellten mittleren Trennschärfekoeffizienten in herkömmlicher und multipler Form für die drei vorgegebenen Untertests ausgehen.

Untertest	Trennschärfe		z-Wert
	herkömmlich	multipel	
Satzergänzung	.36	.46	1.56
Wortauswahl	.28	.37	1.30
Analogien	.38	.45	1.08

Tab. 17: Durchschnittliche herkömmliche und multiple Trennschärfekoeffizienten sowie z -Werte zur Unterschiedsprüfung für die drei IST-Untertests in der Versuchsgruppe (2)

Der Unterschied zwischen herkömmlicher und multipler Trennschärfe liegt im Durchschnitt zwischen .07 und .10 und kann damit nicht als überzufällig angesehen werden. Wir müssen damit unsere entsprechende Hypothese zurückweisen.

c. Vergleich zwischen den beiden Versuchsgruppen

Zunächst gilt es festzustellen, ob sich aufgrund der erweiterten Antwortmöglichkeit in der zweiten Versuchsgruppe der Aufgabenkennwert Trennschärfe wesentlich von dem in der ersten Versuchsgruppe festgestellten unterscheidet. Wir überprüfen dazu die Differenzen zwischen den herkömmlichen Trennschärfekoeffizienten (Korrelation Antwort/Gesamtwert) zwischen den beiden Versuchsgruppen - s.Tab. 18.

Item	Satzergänzung	Wortauswahl	Analogien
1	.22	-[1)]	-
2	.16	.19	-
3	.09	.02	-
4	.07	.22	.12
5	.71*	.02	.35
6	.02	.33	.74*
7	.34	.24	.96**
8	.47	.14	.34
9	.44	.25	.35
10	.20	.25	.35
11	.54	.11	.50
12	.25	.17	.38
13	.31	.10	.15
14	.58	.13	.17
15	.18	.21	.30
16	.12	.45	.16
17	.33	.39	.68*
18	.03	.03	.44
19	.06	.07	.13
20	.32	.32	.07

Tab. 18: Differenzen (absoluter Betrag) zwischen den z'-Werten der Trennschärfekoeffizienten (Item-Rohwert-Korrelation) zwischen Versuchsgruppe (1) und Versuchsgruppe (2) für die drei IST-Untertests

1) Bei diesen Items war keine Differenzbildung zwischen den Trennschärfekoeffizienten aus beiden Versuchsgruppen möglich, da in einer oder beiden Gruppen sie nicht berechnet werden konnten (die Aufgaben wurden von allen Probanden richtig gelöst).

Beim Untertest SE ergibt sich zwischen dem nicht-signifikanten Trennschärfekoeffizienten in Versuchsgruppe (1) und dem auf 1%-Niveau gesicherten in Versuchsgruppe (2) ein statistisch bedeutsamer Unterschied bei der Testaufgabe 5.

Die Versuchsgruppen (1) und (2) unterscheiden sich in den Trennschärfekoeffizienten des Untertests WA statistisch nicht bedeutsam voneinander.

Die auf dem 5%-Niveau (Item 6 und 17) bzw. 1%-Niveau gesicherten (Item 7) Unterschiede zwischen beiden Versuchsgruppen beim Untertest AN gehen in allen drei Fällen von nicht signifikanten negativen Korrelationen bei der Versuchsgruppe (1) aus. Dabei sind die positiven Koeffizienten in der zweiten Versuchsgruppe bei den Items 6 und 7 statistisch bedeutsam, während der Trennschärfekoeffizient der 17. Testaufgabe als bedeutsamer Koeffizient nicht gesichert ist.

Zusammenfassend können wir für die Trennschärfekoeffizienten (Item-Rohwert-Korrelation) feststellen, dass insgesamt kein Unterschied in der Höhe der Koeffizienten zwischen den beiden Antwortformen besteht. Wenn im Einzelfall ein Unterschied besteht, so ist es sehr wahrscheinlich, dass die Testaufgaben beim üblichen Antwortmodus eine nicht signifikante negative Korrelation mit dem Gesamttestwert aufweisen, während bei der Beantwortung mit Angabe der subjektiven Sicherheit bedeutsame positive Trennschärfekoeffizienten vorliegen.

Damit ist unsere Hypothese, dass aufgrund des Antwortmodus in der zweiten Versuchsgruppe bedeutsam höhere Trennschärfekoeffizienten erzielt werden, zurückzuweisen.

Es ist nun zu fragen, ob zwischen den multiplen Trennschärfekoeffizienten, die zusätzlich die zur Bearbeitung benötigte Zeit berücksichtigen, beim Vergleich beider Versuchs-

gruppen bedeutsame Unterschiede festzustellen sind; dazu prüfen wir statistisch die Differenz zwischen den entsprechenden z'-Werten (vgl. Tab. 19).

Item	Satzergänzung	Wortauswahl	Analogien
1	.33	-[1)]	-
2	.09	.16	-
3	.13	.07	-
4	.14	.24	.08
5	.64	.03	.29
6	.11	.11	.21
7	.28	.03	.24
8	.48	.24	.08
9	.02	.13	.23
10	.15	.20	.34
11	.57	.02	.46
12	.34	.17	.11
13	.30	.15	.17
14	.21	.19	.13
15	.12	.24	.04
16	.02	.38	.19
17	.56	.45	.17
18	.36	.15	.12
19	.23	.22	.06
20	.15	.19	.0

Tab. 19: Differenzen (Absoluter Betrag) zwischen den z'-Werten der multiplen Trennschärfekoeffizienten (Item-Rohwert-Korrelation unter Einbeziehung der zur Bearbeitung benötigten Zeit) zwischen Versuchsgruppe (1) und Versuchsgruppe (2) in den drei IST-Untertests.

---

1) Bei diesen Koeffizienten war keine Differenzbildung zwischen den Versuchsgruppen möglich, da sie in einer oder beiden Gruppen nicht berechnet werden konnten (die Aufgaben wurden von allen Probanden richtig gelöst).

Wie Tab. 19 zeigt ergibt sich bei keinem Item der drei Untertests im Vergleich der multiplen Trennschärfekoeffizienten aus beiden Versuchsgruppen ein statistisch bedeutsamer Unterschied; wir müssen deshalb unsere entsprechende Hypothese zurückweisen.

Abschliessend ist zu fragen, ob sich zwischen den durchschnittlichen herkömmlichen bzw. multiplen Trennschärfekoeffizienten aus beiden Versuchsgruppen Unterschiede ergeben; wir gehen dazu von Tab. 20 aus.

Untertest	herkömmlich			multipel		
	VG (1)	VG (2)	z -Wert	VG (1)	VG (2)	z -Wert
Satzergänzung	.29	.36	1.02	.46	.46	.0
Wortauswahl	.29	.28	.14	.32	.37	.72
Analogien	.36	.38	.28	.43	.45	.31

Tab. 20: Mittlere herkömmliche und multiple Trennschärfekoeffizienten sowie z-Werte zur Unterschiedsprüfung in den drei IST-Untertests für die Versuchsgruppen (1) und (2)

Ein Vergleich der mittleren herkömmlichen bzw. mittleren multiplen Trennschärfekoeffizienten in den drei Untertests jeweils zwischen beiden Versuchsgruppen ergibt, dass aufgrund des unterschiedlichen Antwortmodus zwischen beiden Versuchsgruppen weder bei der Item-Rohwert-Korrelation noch bei den Koeffizienten mit Einbezug der Bearbeitungszeit bedeutsame Unterschiede auftreten. Die Art des Antwortmodus trägt also jeweils nicht zur Erhöhung der durchschnittlichen Trennschärfe bei.

Zusammenfassend können wir für den Bereich der Trennschärfe feststellen, dass

o bei herkömmlicher Antwortform die Einbeziehung der Be-

arbeitungszeit nicht durchgängig zur Erhöhung der Trennschärfe beiträgt;

- o bei Antworten mit Angabe der subjektiven Sicherheit die Einbeziehung der Bearbeitungszeit nicht zur Erhöhung der Trennschärfe beiträgt;

- o bei den üblichen Trennschärfekoeffizienten der Antwortmodus keinen Einfluss auf die Höhe der Trennschärfe hat;

- o bei den multiplen Trennschärfekoeffizienten der Antwortmodus keinen Einfluss auf die Höhe der Koeffizienten hat.

C. Innere Konsistenz

Die Leistungsfähigkeit eines Testverfahrens als Messinstrument wird durch die Angabe des entsprechenden Reliabilitätskoeffizienten gekennzeichnet. Dabei kommen Reliabilitätskennwerte, die durch eine Konsistenzanalyse ermittelt werden, der grundlegenden Idee der Reliabilität am nächsten (vgl. LIENERT, 1969). Denn Konsistenzkoeffizienten ermöglichen ein Urteil darüber, innerhalb welcher Grenzen die Messfehler von Testergebnissen liegen.

Zur Abschätzung der Messgenauigkeit (r_{tt}) der einzelnen IST-Untertests bei automatisierter Durchführung gehen wir von der von GULLIKSEN (1950) abgeleiteten Formel aus, die der "KUDER-RICHARDSON-Formula 20" entspricht (LIENERT, 1969):

$$r_{tt} = \frac{n}{n-1} \cdot \left(1 - \frac{\Sigma pq}{(\Sigma r_{it}\sqrt{pq})^2}\right)$$

Dabei bedeuten:

n : Anzahl der Testaufgaben;

p : Schwierigkeitsindex ausgedrückt als Verhältniszahl P / 100 die Wahrscheinlichkeit für eine richtige Lösung;

q : (1 - p), die Wahrscheinlichkeit für eine falsche Lösung;

r_{it} : Trennschärfekoeffizient.

Diese Formel gibt wie "Formula 20" von KUDER-RICHARDSON eine Minimalschätzung der Konsistenz. In sie gehen die Trennschärfekoeffizienten (r_{it}) der Items und die Wahrscheinlichkeiten für die richtige (p) bzw. falsche (q) Lösung der Items ein.

Zur Verdeutlichung der Aussagen über die Messgenauigkeit der Testverfahren ziehen wir im folgenden auch den über r_{tt} definierten Standardmessfehler (s_e) heran, der den Anteil an der Standardabweichung (s_x) eines Tests angibt, der zu Lasten seiner Unreliabilität geht (LIENERT, 1969):

$$s_e = s_x \sqrt{1 - r_{tt}}$$

Dabei bedeuten:

s_e : Standardmessfehler des Tests,

s_x : Standarabweichung des Tests.

Um die Bedeutung des Standardmessfehlers noch anschaulicher darstellen zu können, nehmen wir zusätzlich an, dass die Rohwerte in Z-Werte (Streuung: 10) transformiert seien; somit ergibt sich:

$$s_e = 10 \sqrt{1 - r_{tt}}$$

Wenn im folgenden der Standardmessfehler zahlenmässig dargestellt wird, so ist er jeweils auf Z-Werte bezogen.

Zum Vergleich zwischen den Konsistenzkoeffizienten bzw. Standardmessfehlern, die ohne bzw. mit Einbezug der zur Bearbeitung der Testaufgaben benötigten Zeit errechnet wurden, gehen wir von der o.a. GULLIKSEN-Formel aus. Dazu setzen wir ohne Berücksichtigung der Bearbeitungszeit die im vorhergehenden Abschnitt angeführten herkömmlichen Trennschärfekoeffizienten und mit Einbezug der Zeit die entsprechenden multiplen Trennschärfekoeffizienten für r_{it} ein.

a. Versuchsgruppe (1)

Für die drei IST-Untertests erhalten wir bei herkömmlichem Antwortmodus folgende Werte (Tab. 21).

Untertest	ohne Zeit		mit Zeit		z-Wert
	r_{tt}	s_e	r_{tt}	s_e	
Satzergänzung	.185	9.03	.766	4.84	2.39*
Wortauswahl	.183	9.04	.561	6.63	1.31
Analogien	.235	8.75	.696	5.51	1.81

Tab. 21: Konsistenzkoeffizienten und Standardmessfehler ohne und mit Berücksichtigung der Latenzzeit sowie z-Werte zur Unterschiedsprüfung der Konsistenzkoeffizienten für die drei IST-Untertests in der Versuchsgruppe (1)

Für den Untertest SE ergibt sich eine wesentliche Erhöhung ($z = 2.39$; $P = 2\%$) der Konsistenzkoeffizienten bzw. Verminderung der Standardmessfehler, wenn man nicht nur von der Item-Rohwert-Korrelation ausgeht, sondern zusätzlich die Latenzzeiten berücksichtigt. Dies gilt tendenzmässig auch für den Untertest AN ($z = 1.81$; $P = 8\%$). Unsere entsprechende Hypothese ist damit teilweise bestätigt.

b. Versuchsgruppe (2)

Wenn mit der Beantwortung der Testaufgaben gleichzeitig

die subjektive Sicherheit über die Richtigkeit der Lösung anzugeben ist, so ergibt sich für die Reliabilität der Testverfahren:

Untertest	ohne Zeit		mit Zeit		z-Wert
	r_{tt}	s_e	r_{tt}	s_e	
Satzergänzung	.345	8.09	.785	4.64	2.03*
Wortauswahl	.195	8.97	.608	6.26	1.47
Analogien	.507	7.07	.743	5.07	1.17

Tab. 22: Konsistenzkoeffizienten und Standardmessfehler ohne und mit Berücksichtigung der Latenzzeit sowie z-Werte zur Unterschiedsprüfung der Konsistenzkoeffizienten für die drei IST-Untertests in der Versuchsgruppe (2)

Bei allen drei Untertests zeigen sich mit Berücksichtigung der Latenzzeiten höhere Konsistenzkoeffizienten bzw. niedrigere Standardmessfehler im Vergleich zu den Werten ohne Einbezug dieser Variablen. Statistisch bedeutsam ist dieser Unterschied jedoch nur für den Untertest SE. Unsere entsprechende Hypothese wird damit nur für diesen Untertest bestätigt.

c. Vergleich der beiden Versuchsgruppen

Wir überprüfen zunächst, ob der Antwortmodus einen Einfluss auf die Messgenauigkeit hat, indem wir die Werte ohne Einbezug der Latenzzeit zwischen den beiden Versuchsgruppen vergleichen (vgl. Tab. 23).

Insgesamt finden wir keine bedeutsame Erhöhung der inneren Konsistenz und eine damit verbundene Minderung des Standardmessfehlers bei Testantworten mit Angabe der subjektiven Einschätzung der Antwortrichtigkeit im Vergleich zur herkömmlichen Aufgabenbeantwortung. Die entsprechende Hypothese

können wir damit nicht als bestätigt ansehen.

Untertest	VG (1) r_{tt}	VG (1) s_e	VG (2) r_{tt}	VG (2) s_e	z-Wert
Satzergänzung	.185	9.03	.345	8.09	.50
Wortauswahl	.183	9.04	.195	8.97	.04
Analogien	.235	8.75	.507	7.07	.93

Tab. 23: Konsistenzkoeffizienten und Standardmessfehler sowie z-Werte zur Unterschiedsprüfung der Konsistenzkoeffizienten ohne Berücksichtigung der Latenzzeit in den drei IST-Untertests für die beiden Versuchsgruppen

Bei zusätzlicher Berücksichtigung der Latenzzeit ergeben sich für die beiden Versuchsgruppen:

Untertest	VG (1) r_{tt}	VG (1) s_e	VG (2) r_{tt}	VG (2) s_e	z-Wert
Satzergänzung	.766	4.84	.785	4.64	.15
Wortauswahl	.561	6.63	.608	6.26	.20
Analogien	.696	5.51	.743	5.07	.28

Tab. 24: Konsistenzkoeffizienten und Standardmessfehler mit Berücksichtigung der Latenzzeit in den drei IST-Untertests für die beiden Versuchsgruppen

Die Ergebnisse aus allen drei Untertests sind für beide Versuchspersonengruppen fast genau übereinstimmend. Es ergibt sich damit, dass unter Beachtung der zur Aufgabenbeantwortung benötigten Zeit, die Art der Aufgabenbeantwortung keinen Einfluss auf die Messgenauigkeit des Verfahrens hat. Wir müssen daher unsere entsprechende Hypothese zurückweisen.

Für den Bereich der inneren Konsistenz automatisiert dargebotener Testverfahren können wir festhalten, dass

- o bei herkömmlicher Aufgabenbeantwortung die Reliabilität durch Einbeziehung der Latenzzeit zum Teil wesentlich erhöht wird;

- o bei Aufgabenbeantwortung mit Angabe der subjektiven Antwortsicherheit die Reliabilität durch Einbeziehung der Latenzzeit zum Teil wesentlich erhöht wird;

- o ohne Beachtung der Latenzzeit die Berücksichtigung der subjektiven Antwortsicherheit keine Erhöhung der Reliabilität erbringt;

- o mit Beachtung der Latenzzeit die Art der Aufgabenbeantwortung (Angabe der richtigen Lösung bzw. Angabe der subjektiven Antwortsicherheit) keinen Einfluss auf die Messgenauigkeit hat.

3. DISKUSSION DER BEFUNDE

Global war mit der 2. Versuchsserie die Frage abzuklären, wieweit die Messgenauigkeit eines herkömmlichen Testverfahrens zur Leistungsmessung dadurch verbessert werden kann, dass es in automatisierter Form dargeboten und die mit der Automatisierung des Verfahrens eröffneten Möglichkeiten zusätzlich ausgenutzt werden. Als zusätzliche Informationsquellen wurden die Zeiten bis zur Beantwortung einer Testaufgabe und die Angabe der subjektiven Sicherheit über die Richtigkeit der Antwort herangezogen.

Für den Bereich der Aufgabenschwierigkeit konnten wir insgesamt (ohne und mit Einbezug der Latenzzeiten) keine Unterschiede in den Ergebnissen zwischen den beiden Versuchsgruppen ausmachen; das bedeutet, dass die Aufgabenbeantwortung mit Angabe der subjektiven Sicherheit über die

Richtigkeit der Antwort, auch unter Berücksichtigung der für die Beantwortung benötigten Zeit, keine zusätzlichen diagnostischen Informationen liefert. Damit hat sich hier die von GREEN (1970) geäusserte Vermutung nicht bestätigt, dass die Latenzzeiten zu einer Gewichtung dichotomer Testantworten hinsichtlich der Antwortunsicherheit der Probanden herangezogen werden können.

Für die weiteren Überlegungen hinsichtlich des Zusammenhanges zwischen Aufgabenbeantwortung und Latenzzeit können wir uns damit auf die herkömmliche Beantwortung von Fragen in "multiple-choice"-Form beschränken.

Entgegen der allgemeinen Erwartung konnten wir für das einzelne Item keinen durchgängigen Zusammenhang in dem Sinn feststellen, dass z.B. eine schnelle Bearbeitungszeit für die Richtigkeit der Aufgabenbeantwortung spricht. Für die Praxis der automatisierten Testdarbietung bedeutet das u.a., dass man für die Beantwortungszeit einzelner Items nicht bestimmte Höchstgrenzen festlegen kann, weil anzunehmen ist, dass nach Ablauf dieser Zeit keine richtige Antwort mehr gegeben wird.

Während wir also bei der Anwendung von Testverfahren bezüglich des einzelnen Items aus der Latenzzeitbestimmung keine zusätzlichen diagnostischen Informationen erhalten können, mag sie jedoch in der Phase der Testkonstruktion eine gewisse Bedeutung erhalten. Denn es konnte gezeigt werden, dass nach einer Phase der Eingewöhnung in automatisierte Testsysteme eine bedeutsame Beziehung zwischen der Schwierigkeit von Testaufgaben und der durchschnittlichen Antwortzeit besteht; je schneller eine Aufgabe beantwortet wird, desto leichter ist sie im allgemeinen.

Wir könnten also bei der Testkonstruktion die Latenzzeiten zur genaueren Bestimmung der Aufgabenschwierigkeit heranziehen.

Die Bestimmung von Trennschärfekoeffizienten hat für die praktische Anwendung von Testverfahren nur eine indirekte Bedeutung, indem trennscharfe Aufgaben zur Erhöhung der Messgenauigkeit eines Tests beitragen. Ein diagnostischer Gewinn aufgrund der bei der automatisierten Testdarbietung zusätzlich anfallenden Informationen hätte also vor allem für die Testkonstruktion eine Bedeutung. Doch wie wir gezeigt haben, ermöglicht weder die differenzierte Aufgabenbeantwortung noch die Messung der Bearbeitungszeiten eine deutliche Erhöhung der Trennschärfekoeffizienten. Die Beachtung dieser Variablen kann also auch in der Phase der Testkonstruktion für das einzelne Item kaum zusätzlichen Gewinn bringen.

Wie sich für die innere Konsistenz der Testverfahren ergab, trägt die differenzierte Aufgabenbeantwortung im Vergleich zur herkömmlichen nicht zur Erhöhung der Messgenauigkeit bei, während die Einbeziehung der Latenzzeiten unabhängig von der Art der Aufgabenbeantwortung diese zum Teil wesentlich verbessert. Wir können hier eine teilweise Überlegenheit der automatisierten Testdurchführung gegenüber der herkömmlichen aufzeigen.

Vergleichen wir nun den diagnostischen Wert der beiden zusätzlichen Informationsmöglichkeiten (Aufgabenbeantwortung mit Angabe der subjektiven Sicherheit und Messung der Latenzzeit) so wird allein aufgrund der differenzierten Beantwortung keine Verbesserung der Messgenauigkeit erzielt. Ein Vorteil ergibt sich durch den Einbezug der Latenzzeiten.

Zwar hat die Latenzzeitmessung für das einzelne Item bei weitem nicht die Bedeutung, die ihr oft zugeschrieben wird, doch kann damit die Messgenauigkeit eines Testverfahrens teilweise erheblich verbessert werden.

Mit Blick auf unsere übergreifende Untersuchungsfrage können wir also feststellen, dass die automatisierte Testdurchführung herkömmlicher Verfahren abgesehen von dem möglichen ökonomischen Gewinn aufgrund der Erhöhung der Messgenauigkeit einen zusätzlichen diagnostischen Gewinn mit sich bringt. Für die Praxis bedeutet dies: wenn ein Verfahren automatisiert durchgeführt wird, so ist auf jeden Fall die Zeit bei der Beantwortung der einzelnen Aufgaben zu messen, um durch eine nach Beantwortungszeiten differenzierte Aufgabenbewertung (z.B. Staffelung der erreichbaren Punkte nach der Lösungszeit) zu reliableren Gesamttestergebnissen zu kommen.

Damit ist das erste Ziel der vorliegenden Arbeit erreicht: Für die in der Praxis üblichen Testverfahren, die nach den Prinzipien der klassischen Testtheorie entwickelt worden sind, konnten wir am Beispiel der Latenzzeitmessung aufzeigen, dass die automatisierte Testdurchführung der herkömmlichen Darbietung zumindest dann überlegen ist, wenn wenigstens einzelne der bei dieser Durchführungsart zusätzlich zu erhaltenen Informationen aufgegriffen werden. Die Einbeziehung weiterer von verschiedenen Seiten vorgeschlagener Informationsquellen kann aufgrund der exemplarischen Überprüfung zweier Variabler in dieser Untersuchung den diagnostischen Wert herkömmlicher Testverfahren bei automatisierter Darbietung nur erhöhen.

NEUE MÖGLICHKEITEN BEI COMPUTERUNTERSTÜTZTEM TESTEN

Im bisherigen Verlauf dieser Arbeit haben wir untersucht, welche Möglichkeiten durch den Computer-Einsatz in der diagnostischen Datenerhebung eröffnet werden; abgesehen von einem eventuellen ökonomischen Vorteil ergibt sich aus der automatisierten Durchführung eines herkömmlichen Testverfahrens aufgrund der extremen Auswertungsobjektivität und vor allem durch die Latenzzeitbestimmung eine Erhöhung der Messgenauigkeit.

Prinzipiell besteht jedoch kein Unterschied zur herkömmlichen Testdurchführung. Der automatisierte "Untersucher" konfrontiert den Probanden mit einem vorliegenden Testverfahren, ist stets gleich "freundlich", hält sich bei der Durchführung genau an die Handanweisung, registriert penibel Zeiten und Antworten und zählt die Anzahl der richtigen Lösungen grundsätzlich fehlerfrei aus.

Will man sich nicht darauf beschränken, den Testleiter durch den Computer zu ersetzen, so bieten sich m.E. zumindest die Bereiche "Testdurchführung" und "Testkonstruktion" dazu an, durch das Aufgreifen der spezifischen Möglichkeiten, die der Computereinsatz mit sich bringt, zu einer genaueren Leistungsmessung zu kommen.

(1) Zum einen heisst es umdenken bei der Testdurchführung: Der grösste Teil der bestehenden Leistungstests ist so angelegt, dass eine genaue Leistungsbestimmung nur im mittleren Bereich möglich ist, während zur Differenzierung im unteren und oberen Bereich nur wenige Items zur Verfügung stehen. Auch wird meist allen Probanden derselbe Itemsatz vorgelegt, von dem ein Teil für den einzelnen Probanden zu leicht, ein anderer zu schwer ist, und daher aus der Beantwortung dieser Testaufgaben keine relevanten Informationen zur Kennzeichnung des Probanden entnommen werden können.

Hier könnte der Computer die für den einzelnen Probanden kritischen Aufgaben sehr schnell auswählen und aufgrund ihrer Beantwortung die Fähigkeit des Probanden bestimmen. Dieses unter dem Namen "tailored testing" bekannte Verfahren (vgl. z.B. LORD, 1970) erfordert aber einen sehr grossen "Itempool" und wird u.a. deshalb in der Praxis bisher kaum angewandt.

(2) Der zweite Grund zum Überdenken bisher festgelegter Vorstellungen ergibt sich eigentlich schon zwangsläufig aus der Notwendigkeit eines grossen "Itempools" beim "massgeschneiderten Testen": Warum soll der Computer nur dazu benutzt werden, eine Reihe vorher konstruierter Items zu speichern und diese aufgrund bestimmter Regeln abzurufen, und nicht auch dazu, die Testaufgaben jeweils neu zu konstruieren?

In der weiteren Darstellung wollen wir nun so vorgehen, dass wir zunächst kurz das "logistische Messmodell" ansprechen, das den theoretischen Hintergrund für die dann zu diskutierenden Punkte des "tailored testing" und der "aufgabenorientierten Schwierigkeitsbestimmung"von Testaufgaben bietet, um danach einen neuen Ansatz zum computerunterstützten Testen zu entwickeln.

I. DIE BISHER ERARBEITETEN GRUNDLAGEN

1. DAS LOGISTISCHE MESSMODELL

Es ist hier nicht der Ort, auf die Kritik an der klassischen Testtheorie einzugehen, die wegen der relativen Schwächen ihrer Annahmen von LORD & NOVICK (1968) als "weiches" Modell bezeichnet wird (zur Kritik vgl. z.B. FISCHER, 1974), sondern wir wollen nur kurz die Grundgedanken der von der klassischen Testtheorie sich absetzenden probabilistischen Testmodelle ansprechen.

Diese Modelle stehen im Gegensatz zur klassischen Testtheorie, indem sie von der Annahme einer "latenten", nicht direkt zu beobachtenden Dimension ϑ (Theta) ausgehen, die nicht mit den beobachtbaren "manifesten" Variablen, den Testwerten im allgemeinen Sinn, gleichzusetzen ist.

"Was wir beobachten, sind die Symptome der latenten Eigenschaften, auf welche man nur indirekt zurückschliessen kann. Der wesentliche Unterschied zur klassischen Testtheorie besteht darin, dass diese die latente Eigenschaft mit beobachtbaren Grössen, wie z.B. der Zahl der von einer Vp in einem Test gemachten Fehler gleichsetzt; da ein solcher Ansatz seiner Natur nach deterministisch ist - latente Dimension ist gleich beobachtetes Verhalten -, ist er der empirischen Realität nicht adäquat und zwingt dazu, gewissermassen nachträglich ein additives Fehlerglied mit Zufallsverteilung einzuführen." (FISCHER, 1974, S.148)

Kennzeichnend für diesen Ansatz, der in unserem Sprachgebiet zunächst als RASCHsches Modell (s.u.a. RASCH, 1960) bekannt geworden ist, sind drei Prinzipien.

- Latente Dimension und beobachtete Variablen werden in dem Sinn getrennt aufgefasst, dass die beobachteten Variablen nur als Symptome oder Indikatoren der latenten Dimension betrachtet werden.

- Der probabilistische Charakter der beobachteten Variablen wird berücksichtigt, indem im Gegensatz zum deterministischen Ansatz der klassischen Testtheorie der Tatsache Rechnung getragen wird, dass das Auftreten von manifesten Variablen nur mit einer bestimmten Unsicherheit vorhergesagt werden kann.

- Es gilt das Prinzip der lokalen stochastischen Unabhängigkeit von Verhaltensweisen; d.h. die Korrelation zwischen beobachteten Verhaltensweisen wird nicht als Abhängigkeit zwischen den Verhaltensweisen selbst angesehen, sondern als Abhängigkeit von einer gemeinsamen latenten Dimension.

Die Forderung nach der stochastischen Unabhängigkeit von Verhaltensweisen ist bei den Aufgaben eines Tests nur dann erfüllt, wenn die Wahrscheinlichkeit, alle Aufgaben richtig zu beantworten, sich multiplikativ aus den Lösungswahrscheinlichkeiten für die einzelnen Items bestimmen lässt.

Die Grundgleichung des logistischen Messmodells besagt, dass die Wahrscheinlichkeit für eine richtige bzw. falsche Antwort, eindeutig durch den Personenparameter ξ (Xi) und den Aufgabenparameter ε (Epsilon) gegeben ist, sie wird jedoch nicht davon bedingt, ob und welche Aufgaben dieses oder eines anderen Tests der Proband vorher schon gelöst hat.

Diese Grundgleichung lässt sich schreiben (RASCH, 1966) als:

$$p\left[\left(\begin{smallmatrix} g \\ + \end{smallmatrix}\right)\middle| v\right] = \frac{\xi_v \cdot \varepsilon_g}{1 + \xi_v \cdot \varepsilon_g}$$

Dabei bedeuten:

ξ_v : die Fähigkeit der Person v;

ε_g : die Leichtigkeit der Aufgabe g;

$p\left[\binom{g}{+} \mid v\right]$: die bedingt Wahrscheinlichkeit, dass ein Proband v mit der Fähigkeit ξ_v die Aufgabe g mit der Leichtigkeit ε_g lösen kann.

Für die weitere Darstellung benutzen wir nun die stärker verbreitete Notation von LORD & NODIK (1968); dabei ist:

$$\vartheta_v = \ln \xi_v ;$$

$$b_g = \ln \frac{1}{\varepsilon_g} ;$$

$$p\left[\binom{g}{+} \mid v\right] = p_g(\vartheta) = p_g$$

Formelt man die Grundgleichung von RASCH unter Benutzung dieser Notation etwas um, so lässt sich der logistische Charakter dieses Modells aufzeigen (vgl. FRICKE, 1972):

$$p_g = \frac{1}{1 + \exp(-(\vartheta - b_g))}$$

Die Verteilungsfunktion der logistischen Funktion hat fast dieselbe Form:

$$y = \frac{1}{1 + \exp(-Da(x-b))}$$

Die Verteilungsfunktion der logistischen Funktion entspricht der Verteilungsfunktion einer Normalverteilung mit dem Mittelwert 0 und einer Standardabweichung von 1,7.

Setzt man für D in dieser Funktion 1,7 ein, so ergibt sich eine maximale Übereinstimmung der Werte mit denen der Normalverteilung (LORD & NOVICK, 1968, S. 399). Dabei gibt a die Steigung der Funktion und b die Verschiebung parallel zur Abzisse an.

Vergleichen wir diese Funktion mit der umgeformten RASCH-Gleichung, so zeigt sich, dass eine Aussage getroffen wird über den Zusammenhang zwischen der zugrunde liegenden Fähigkeit und der Wahrscheinlichkeit, die Aufgabe mit dem Schwierigkeitsgrad b_g zu lösen. Der Faktor a, den man als Trennschärfe zwischen Fähigkeit und Aufgabe auffassen kann, geht dabei in das RASCH-Modell nicht ein; dieses fordert vielmehr für alle Items eine Trennschärfe von 1.

Die Trennschärfe eines Items - wobei zu beachten ist, dass es sich nicht um die aus der klassischen Testtheorie gewohnte Korrelation zwischen Itembeantwortung und Gesamttest-Rohwert, sondern um die Beziehung zwischen Itembeantwortung und zugrunde liegender Fähigkeit handelt, geht in das logistische Modell von LORD & NOVICK (1968) ein. Es lautet dann:

$$p_i(\vartheta) = \frac{1}{1 + \exp(-1.7a_i(\vartheta - b_i))}$$

wobei a_i die Diskriminationsfähigkeit des Items i bedeutet.

Diese Funktion wurde von BIRNBAUM (1968) zu einer logistischen Funktion mit drei Parametern erweitert, um auch solche Items erfassen zu können, die durch Raten lösbar sind. Wenn c_i die Wahrscheinlichkeit angibt, Item i durch Raten richtig lösen zu können, ergibt sich:

$$p_i(\vartheta) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\vartheta - b_i))}$$

Eine "item characteristic curve" (ICC) stellt dann die Wahrscheinlichkeit der richtigen Beantwortung eines Items als Funktion der zu messenden Fähigkeit graphisch dar -vgl. Abb. 17.

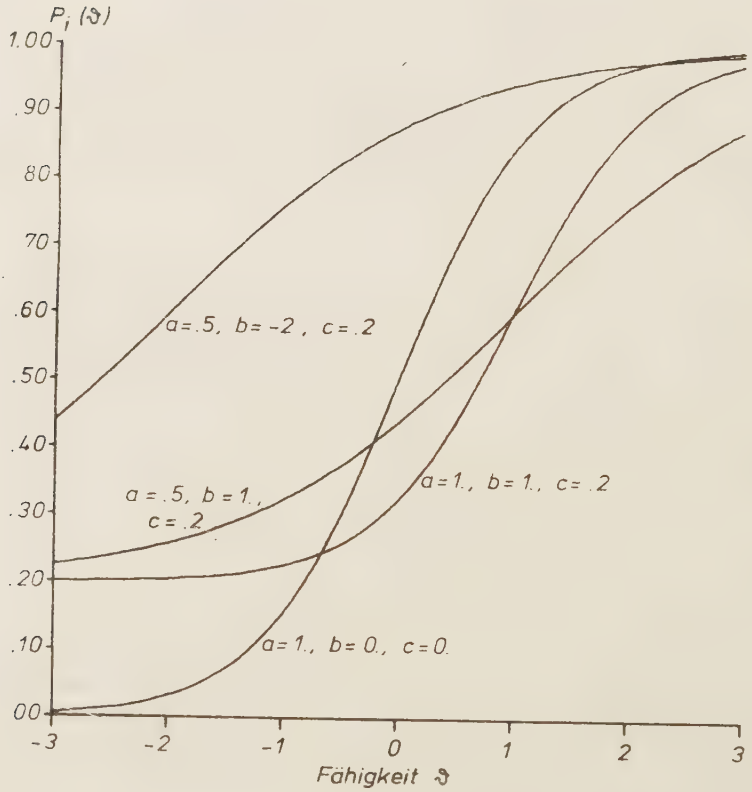


Abb. 17: ICCs mit unterschiedlichen Werten für die Parameter a,b und c.

In dieser Abbildung 17 zeigt z.B. die Kurve mit den Parametern (a=1, b=0, c=0) ein Item nach dem RASCH-Modell mit mittlerer Schwierigkeit; da in dem RASCH-Modell für a immer 1 gefordert wird, würden die ICCs für Items mit anderer Schwierigkeit b lediglich jeweils nach links oder rechts parallel verschoben sein. Der Parameter c - die Ratewahrscheinlichkeit - hat keinen Einfluss auf die Form der Kurve, sondern diese wird mit höherer Ratewahrscheinlichkeit immer weiter vertikal verschoben; die Form der Kurve ist also allein vom Diskriminierungsparameter a abhängig.

Wenn eine Aufgabe nur mit 0 oder 1 verrechnet wird, stellen die ICCs gleichzeitig die Regressionen zwischen dem Item und der Fähigkeit ϑ dar. So kann aufgrund dieser Kurven die Wahrscheinlichkeit bestimmt werden, mit der ein Proband, für den bestimmtes Mass der Fähigkeit geschätzt ist, ein Item richtig beantworten wird. Damit ist eine wichtige Voraussetzung für das "tailored Testing" geschaffen: es können immer solche Items ausgewählt werden, die mit etwa 50-prozentiger Wahrscheinlichkeit gelöst werden können.

Es würde den Rahmen dieser Arbeit überschreiten, hier auf die verschiedenen Methode zur Schätzung der Aufgabenparameter in logistischen Modellen einzugehen. - Es muss auf LORD & NOVICK (1968), LORD (1970.1971) und FISCHER (1974) verwiesen werden.

2. TAILORED TESTING

Die Leistung eines Probanden inbezug auf irgendeine "Fähigkeit" lässt sich dann am besten messen, wenn die vorgegebenen Testaufgaben für den Probanden weder zuschwer noch zu leicht sind. Zur Verwirklichung dieses Vorgehens ist es

notwendig, nach jedem Schritt der Testdurchführung aufgrund der bisher bearbeiteten Items die "Fähigkeit" des Probanden in etwa abschätzen zu können. Da solche Schätzungen und die Entscheidung darüber, welches Item als nächstes vorzugeben ist, nur von einem Computer schnell genug vorgenommen werden können, haben sich für diese Art des Testens die Begriffe "branched testing", "programmed testing", "sequential item testing" bzw. "computerized testing" eingebürgert. LORD (1970 benutzt den Begriff "tailored testing", um nicht so sehr die Art dieses Vorgehens sondern den Grundgedanken stärker zu kennzeichnen.

Wenn wir einen Probanden bezüglich einer Fähigkeit messen wollen, so setzen wir damit voraus, dass es irgendeinen nummerischen Wert ϑ gibt, der diese Fähigkeit kennzeichnet. Zur Bestimmung dieses Wertes ziehen wir die Antworten des Probanden auf bestimmte Testaufgaben heran. Dabei ergibt sich das Problem, eine bestimmte Anzahl von Items so auszuwählen, dass wir ϑ aufgrund ihrer Beantwortung möglichst gut schätzen können.

LORD (1970, 1971), dessen Darlegungen wir hier weitgehend folgen, führt aus, dass daher die optimale Zusammenstellung der n Items für die Messung des Probanden von dem unbekannten Wert ϑ abhängig ist. - Jedoch abgesehen davon, dass unbekannt ist, kann auch nicht als sicher vorausgesetzt werden, dass es überhaupt eine optimale Strategie zur Auswahl der benötigten Items gibt. - Mit einer Reihe möglicher Strategien wollen wir uns jetzt beschäftigen.

A. Strategien zur Itemauswahl

Wir gehen bei unseren Überlegungen davon aus, dass dem Probanden jedes Item einzeln dargeboten wird. Nach der Vorgabe

der ersten Aufgabe soll nun jedes folgende Item nach einer vorher festgelegten Regie ausgewählt werden, wobei diese Entscheidung allein davon abhängt, wie der Proband das vorhergehende Item beantwortet hat. Welches Item dann ausgewählt wird, hängt von der gerade benötigten Aufgabenschwierigkeit ab. Allgemein wird man so vorgehen, dass die Schwierigkeit des neuen Items nach richtiger Lösung des vorhergehenden grösser, nach falscher kleiner als die des vorgehenden ist. LORD (1970, 1971) diskutiert eine Reihe möglicher Vorgehensweisen bei der Itemauswahl, die sich grob in die Gruppe der Strategien mit vorher festgelegtem Schwierigkeitsunterschied "d" und in die Gruppe mit verändertem "d" im Laufe des Tests unterteilen lassen.

Für das Vorgehen mit festgesetztem Schwierigkeitsunterschied lässt sich zeigen (vgl. LORD, 1971), dass für Multiple-choice-Aufgaben mit fünf Antwortmöglichkeiten eine geringe Schrittgrösse von d= .167/a[1] die besten Ergebnisse brachte. Dabei wird bei richtiger Antwort die Schwierigkeit um 1d erhöht, bei einem Fehler um 2d erniedrigt. Mit diesem asymetrischen Vorgehen, sollen richtige Lösungen, die durch Raten entstanden sind, ausgeglichen werden.

Bei einem Vergleich mit dem sogenannten "Standard-Test" - das ist ein Test, dessen Items für eine homogene Probandenstichprobe nur mittlere Schwierigkeit besitzen - differenziert der "tailored test" im oberen und unteren Drittel besser, im mittleren Drittel schlechter als der"Standard-Test".

Vorgehensweisen mit kleiner werdender Schrittgrösse werden allgemein als ROBBINS-MONRO-Prozesse bezeichnet. Solche Prozesse erlauben im Anfang der Untersuchung, wenn die Fähig-

---

[1] "a" gibt die Diskriminierungsfähigkeit des Items bezüglich ϑ an.

keit des Probanden nur grob geschätzt werden kann, ein schnelles Anpassen der Aufgabenschwierigkeit an die Fähigkeit des Probanden und eine immer genauere Schätzung der angemessenen Aufgabenschwierigkeit mit steigender Genauigkeit der Fähigkeitsschätzung. LORD (1971) erzielte die besten Ergebnisse mit folgenden Werten:

Die Schwierigkeit des auf das Item i folgende Item i+1

$$b_{i+1} = b_i + d_i \ (u_i - .5)$$

Dabei ist $u_i = 0$, wenn Item i falsch gelöst wurde und $u_i = 1$ bei richtiger Lösung. Der Schwierigkeitsunterschied d_i wird bestimmt über die harmonische Zahlenreihe:

$$d_i = d_1 / i \quad \text{für: } i = 2,3, \ldots.$$

Die besten Lösungen erbrachte dieses Vorgehen wenn die anfängliche Schrittgrösse $d_1 = 2.5/a$ bzw. $3/a$ gewählt wurde. LORD diskutiert in diesem Zusammenhang eine Reihe von Untersuchungen mit wechselnden Werten für d_1, auf die hier jedoch nicht näher eingegangen werden soll.

Undiskutiert muss auch die Frage nach der Bestimmung der Fähigkeit des Probanden nach Abschluss eines "Tailored Test" bleiben, da sie von der gewählten Stratgie abhängig ist. Nähere Einzelheiten finden sich wiederum bei LORD (1970, 1971) und auch bei GREEN (1970).

B. Vor- und Nachteile des Verfahrens

Das grösste Hindernis, "Tailored Tests" in der Praxis anzuwenden, liegt in der Grösse des benötigten Itempools. Wir wollen das zunächst an einem hypothetischen Beispiel (vgl. WATERS & BAYROFF, 1971) zeigen. Nehmen wir den einfachsten

Fall an, dass bei fester Schrittweite, die Lösungswahrscheinlichkeit des folgenden Items jeweils um .10 erhöht oder vermindert wird. Dann ergeben sich bei nur fünf Items, die jedem Probanden dargeboten werden, folgende Verzweigungen:

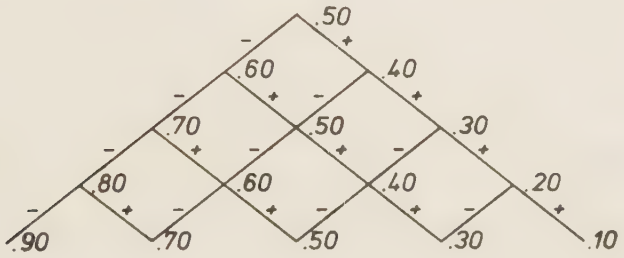


Abb. 18: Verzweigungsmöglichkeiten bei einem 5-Item-tailored-test mit fester Schrittweite.

Allgemein wird bei fester Schrittweite und einer vorgesehenen Testlänge von n Items ein Itempool von $n(n + 1) / 2$ und bei einem ROBBINS-MONRO-Prozess ein Itempool von $2^n - 1$ Items benötigt. Bei einer vorgesehenen Testlänge von 20 Items würden im ersten Fall 210 und im zweiten über eine Million Items benötigt. - Diese Beziehung verdeutlicht Abb. 19.

Wenn ein so grosser Itempool für das "tailored testing" benötigt wird, stellt sich natürlich die Frage, ob dieser Aufwand vertretbar und sinnvoll ist. Geht man von der gewünschten Messgenauigkeit aus, so ist der oben beschriebene optimal "Standard Test" dem Tailored Test" bei Messungen im mittleren Fähigkeitsbereich überlegen, und auch der konventionelle Test erreicht in diesem Bereich fast die Genauig-

keit des "Tailored Tests". Das "massgeschneiderte Verfahren" ist in der Präzision seiner Ergebnisse den beiden anderen Verfahren in den Extrembereichen weit überlegen; hier wäre es also, wenn die Kostenfrage von untergeordneter Bedeutung ist, das Instrument der Wahl.

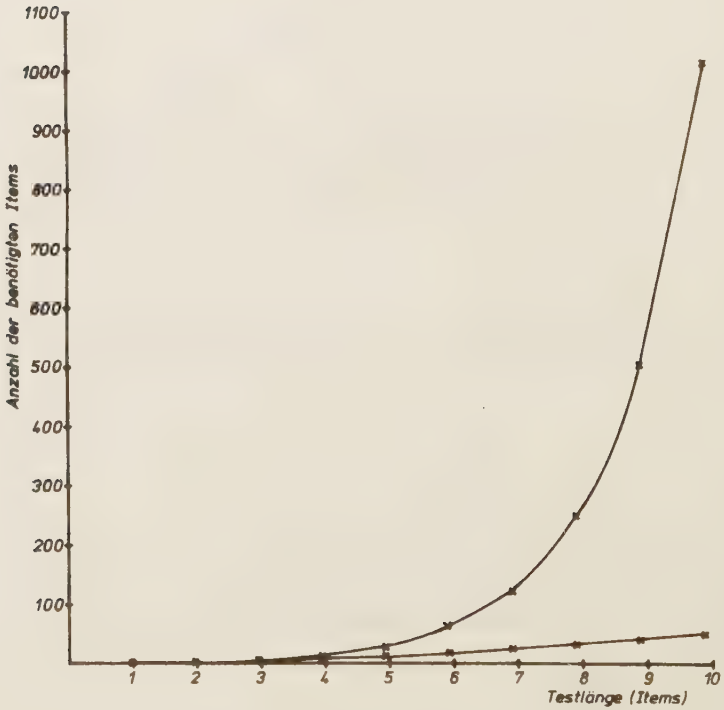


Abb. 19: Grösse des benötigten Itempools beim "tailored testing" in Abhängigkeit von der Testlänge. Untere Kurve: Vorgegeben mit fester Schrittweite; obere Kurve: ROBBINS-MONRO-Prozess.

Beim Vergleich des Tailored Tests mit dem üblichen Test sind wir bisher davon ausgegangen, dass bei beiden Verfahren den Probanden die gleiche Anzahl von Items dargeboten wird. Allgemeiner gesprochen gehen wir in dieser Sicht vom "Messen an sich" aus, und es wird dabei nicht beachtet, wozu die gemessenen Werte benutzt werden sollen. LINN et al. (1969) konnten aufgrund von Simulationsstudien nachweisen, dass ein konventioneller Test etwa 3,4 mal soviele Aufgaben enthalten muss wie ein Tailored Test, um dieselbe mittlere Korrelation mit einem Aussenkriterium zu erreichen.

Von einem ähnlichen Standpunkt aus lässt sich mit GREEN (1970) argumentieren. Dabei gehen wir von der in der Praxis häufig auftretenden Fragestellung aus, ob die Fähigkeiten eines Probanden oberhalb oder unterhalb eines bestimmten Wertes liegen, um ihn aufgrund dieses Wertes einer bestimmten "Leistungsgruppe" zuweisen zu können. Der Vorteil des "tailored testing" besteht dann darin, dass der Test abgebrochen werden kann, sobald mit vorher bestimmter Genauigkeit entschieden werden kann, zu welcher der beiden Gruppen der Proband gehört.

GREEN berechnet aufgrund der Methoden der STATISTICAL RESEARCH GROUP OF COLUMBIA UNIVERSITY (1945), die Anzahl der beim "tailored testing" vorzugebenden Items, die in Abhängigkeit von der Fähigkeit des jeweiligen Probanden dieselbe Aussage ermöglichen, wie ein konventioneller Test mit einer bestimmten feststehenden Anzahl von Items. Diese Beziehung ist in Abb. 20 dargestellt.

Die Anzahl der notwendigen Items ist natürlich von Proband zu Proband unterschiedlich, doch zeigt sich, das selbst im mittleren Fähigkeitsbereich der Tailored Test im Durchschnitt weniger Items benötigt, und in den Extrembereichen wesentlich kürzer wird. Wenn man weiterhin bedenkt, dass im

allgemeinen der Psychologe eine Entscheidung nicht aufgrund der Messung einer Fähigkeit trifft, werden die Vorteile des Tailored Tests noch offenkundiger; die Zahl der bis zu einer Entscheidung zu erhebenden Informationen könnte erheblich verringert werden.

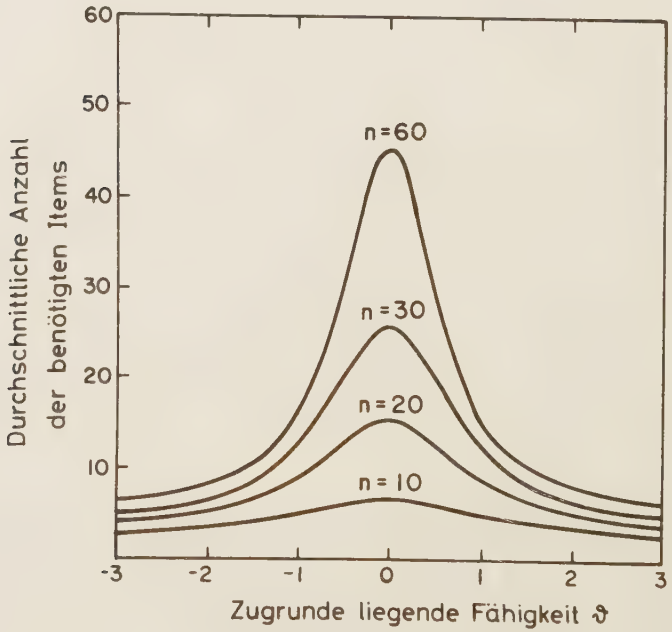


Abb. 20: Durchschnittliche Anzahl der beim "tailored testing" in Abhängigkeit vom Fälligkeitsgrad benötigten Items im Vergleich zu einem herkömmlichen Test mit bestimmter Testlänge n (vgl. GREEN, 1970)

3. AUFGABENORIENTIERTE SCHWIERIGKEITSSCHÄTZUNG

Wie wie gezeigt haben, erfordert das "tailored testing" eine so grosse Menge von Testaufgaben unterschiedlichster Schwierigkeitsgrade, dass dieses Vorgehen bis heute noch keine praktische Bedeutung erlangt hat. Würde man jedoch über explizite Regeln verfügen, die es gestatten, Schwierigkeitsunterschiede zwischen Testaufgaben nicht aufgrund ihrer späteren Bearbeitung in der Analysestichprobe, sondern von vornherein aufgrund ihrer Aufgabenstruktur abzuschätzen, so liessen sich prinzipiell Algorithmen zur Konstruktion von Aufgaben mit vorgegebenen Schwierigkeitsunterschieden entwerfen.

Wir wollen im folgenden auf verschiedene Ansätze zur "aufgabenorientierten" Schwierigkeitsschätzung - "aufgabenorientiert" um den festgelegten Begriff "populationsunabhängig" zu vermeiden - eingehen. Damit soll verdeutlicht werden, dass in unterschiedlichen psychologischen Disziplinen Versuche unternommen werden, Konstruktionsregeln für Aufgaben mit vorgegebenem Schwierigkeitsgrad zu entwickeln. Doch wird dort die Bedeutung dieser Ansätze für die praktische Diagnostik und besonders der Bezug zum "tailored testing" nicht gesehen.

Dabei ist anzumerken, dass der Anstoss zur expliziten Formulierung von Regeln zur aufgabenorientierten Schwierigkeitsschätzung nicht aus der Testkonstruktion, sondern aufgrund von Arbeiten zur CAI kam. Wie UTTAL (1967) feststellte, würde ein grosser Fortschritt für die CAI darin liegen, diejenigen Systeme, die nur vorgefertigte Aufgaben speichern, durch solche zu ersetzen, die nach festgelegten Regeln Aufgaben generieren können.

A. Erste Ansätze im Bereich der CAI

Das oben erwähnte Mathematik-Lehrprogramm COACH (WEXLER, 1970) enthält ein Unterprogramm zur Erzeugung von Rechenaufgaben mit vorgegebener Schwierigkeit. Da WEXLER in der Literatur keine verwertbaren Hinweise fand, entwickelte er eine Methode zur Schätzung der Aufgabenschwierigkeit (problem difficulty score), die sich aufgrund der Bearbeitungsschwierigkeiten (operand difficulty score) der in der Aufgabe enthaltenen Zahlen ergibt. Die Bearbeitungsschwierigkeit errechnet sich aus der Schwierigkeit der einzelnen Ziffern einer Zahl (digit score) und der Stellung der Ziffern in der Zahl (position score). Im einzelnen umfasst die Methode von WEXLER folgende Schritte:

- Ermittlung des "digit score"; da einige Ziffern leichter, andere schwerer zu bearbeiten sind, werden den einzelnen Ziffern unterschiedliche Gewichte zugeteilt.

- Ermittlung des "position score"; je nach Stellung in der Zahl enthalten die Ziffern unterschiedliche Gewichte.

- Ermittlung des "operand difficulty score"; die Bearbeitungsschwierigkeit einer Zahl ist die Summe der Produkte aus "digit score" mal "position score" jeder Ziffer einer Zahl.

- Ermittlung des "problem difficulty score"; die Schwierigkeit einer Aufgabe setzt sich zusammen aus den Bearbeitungsschwierigkeiten der in der Aufgabe vorkommenden Zahlen, die jeweils mit einem Faktor für die Rechoperation gewichtet werden.

Während WEXLER die Schwierigkeit einer Aufgabe über die vorkommenden Rechenoperationen und die zu bearbeitenden

Zahlen bestimmt, gehen SUPPES, JERMAN &BRIAN (1968) für das 'Stanford Arithmetic Program' von der Art der in der Aufgabe durchzuführenden Schritte aus. Ihr Modell für die Berechnung der Kennwerte für Rechenaufgaben wird hier am Beispiel der Addition vorgestellt.

Die Autoren schätzen die Schwierigkeit einer Additionsaufgabe über die vermutete Anzahl der zur Lösung der Aufgabe erforderlichen Schritte; diese Variable NSTEPS setzt sich additiv aus den Anzahlen der Transformations-, Operations- und Memoryschritte zusammen.

- "Transformation" bedeutet die Umformung einer Aufgabe in die sogenannte kanonische Form; d.h. die zu berechnende Unbekannte steht allein auf der rechten Seite des Gleichheitszeichens.

- "Operation" wird dann gezählt, wenn zwei einzelne Ziffern addiert werden müssen, wobei die Addition der Null nicht mitgezählt wird.

- "Memory" ist ein Mass für die Anzahl der Ziffern oder Summen, die während der Rechnung im Gedächtnis behalten werden müssen.

Für die Aufgabe 25 + 26 = 18 + ? z.B. ergeben sich nach dieser Regeln 1 Transformationsschritt, 6 Operationsschritte und 7 Memoryschritte; NSTEPS erhält damit den Wert 14.

SUPPES, JERMAN & BRIAN berechneten dann multiple Korrelationen zwischen NSTEPS und weiteren formalen Aufgabenwerten und der transformierten Leichtigkeitsindizes z_i der Aufgaben. Dabei ist

$$z_i = \ln ((1 - P_i) / P_i)$$

hier bezeichnet P_i die Lösungswahrscheinlichkeit einer Aufgabe, die aufgrund des Lösungsprozentsatzes des Items i in der Stichprobe geschätzt wird.

In allen Regressionsanalysen ergab sich für die Variable NSTEPS das stärkste Gewicht, während die anderen Kennwerte vernachlässigt werden konnten. Bei der Aufsplitterung von NSTEPS in die drei additiven Komponenten Transformation (T), Operation (O) und Memory (M) und die Berechnung der Regression

$$z_i = b_1 T_i + b_2 O_i + b_3 M_i + b_c$$

ergab sich ein multipler Korrelationskoeffizient zwischen z und den drei Aufgabenkennwerten von .731. Das bedeutet, dass über diese Kennwerte die populationsabhängige Schwierigkeit einer Aufgabe recht gut vorhergesagt werden kann. In weiteren Berechnungen stellte sich heraus, dass der Variablen "Memory" die stärkste Bedeutung zukommt; die einfache Korrelation zwischen "Memory" und dem transformierten Leichtigkeitsindex ergab einen Koeffizienten von .726.

Die Ergebnisse von SUPPES, JERMAN & BRIAN konnten von FRICKE (1972) bestätigt werden. FRICKE stellt dem Modell von SUPPES, JERMAN & BRIAN, das auf die Variable Memory bezogen lautet

$$\ln ((1 - P_i) / P_i) = z_i = \gamma_1 \, \mathrm{MEMORY}_i + \gamma_o$$

die in dem RASCH-Modell definierte Lösungswahrscheinlichkeit einer Aufgabe gegenüber:

$$\ln ((1 - P_{ji}) / P_{ji}) = z_{ji} = b_i - \vartheta_j$$

Dabei ist P_{ji} die Wahrscheinlichkeit, dass Personen aus der Fähigkeitsgruppe j (Rohwertgruppe) die Aufgabe i lösen.

Bezieht man das RASCH-Modell auf Lösungswahrscheinlichkeiten, die nicht für einzelne Fähigkeitsgruppen, sondern für die Gesamtgruppe gelten, so ergibt sich:

$$\ln ((1 - p_i) / p_i) = z_i = b_i - \bar{\vartheta}$$

Dabei entspricht $\bar{\vartheta}$ dem arithmetischen Mittel der Fähigkeitsparameter in den verschiedenen Rohwertgruppen.

"Die Variable γ_1 MEMORY des SUPPES-Modells ist also mit dem Schwierigkeitsparameter b des RASCH-Modells und die Variable $-\gamma_0$ mit dem arithmetischen Mittel aller Fähigkeitsparameter $\bar{\vartheta}$ der Testwertgruppen gleichzusetzen, wenn beide Modelle die z_i-Werte gleichermassen gut vorhersagen. Das SUPPES-Modell entspricht also einem RASCH-Modell, das keine individuellen Fähigkeitsunterschiede zulässt." (FRICKE, 1972, S. 102).

Zur Überprüfung der Schätzung der Schwierigkeitsparameter über die von SUPPES, JERMAN & BRIAN entwickelten formalen Aufgabenwerte wählte FRICKE aus dem diagnostischen Rechentest "Zahlenrechnen ZR4+" für 4. und höhere Klassen (KRÜGER, HYLLA, BARGMANN & HORN, 1965) den Untertest mit den meisten Additions- und Subtraktionsaufgaben (1. Untertest mit den Aufgaben 1 - 14) aus. Für diese Aufgaben wurden die Schwierigkeitsparameter b nach dem RASCH-Modell geschätzt, wobei jedoch die Aufgaben 1 - 4 aufgrund der Modellbedingungen eliminiert werden mussten.

Die Korrelation zwischen der SUPPES-Variable NSTEPS und dem RASCH-Parameter b ergab einen Koeffizienten von .93. Wurde statt NSTEPS - einem additiven Wert - nur die Variable "Memory" mit dem Parameter b korreliert, so reduzierte sich

die Korrelation lediglich auf .90. "Solange wir also keine anderen Korrelationen zwischen den manifesten Aufgabenkennwerten und den latenten Aufgabenparametern besitzen, können wir sagen, dass das hypothetische Konstrukt "Rechenschwierigkeit' durch die Zahl der zu merkenden Zwischenergebnisse beschrieben werden kann"(FRICKE, S. 100).

Auch wenn FRICKE die Schwierigkeitsschätzungen für Rechenaufgaben im Nachherein an einem bestehenden Test vornahm, können wir trotzdem festhalten, dass die von SUPPES et al. entwickelten Kriterien zur Vorhersage von Aufgabenschwierigkeiten brauchbar sind.

Nun kann eingewandt werden, dass es sich bei Rechenaufgaben um Items handelt, deren Schwierigkeit am ehesten formal bestimmt werden kann, doch sollte das kein Hinderungsgrund sein, diese Ansätze auch auf andere inhaltliche Bereiche zu übertragen; auf solche Arbeiten ist im nächsten Abschnitt näher einzugehen.

B. Aufgabenkonstruktion nach dem logistischen Modell

Während die Modelle von WEXLER bzw. SUPPES, JERMAN & BRIAN aus der praktischen Notwendigkeit einer automatischen Itemkonstruktion entwickelt wurden, ist den folgenden Untersuchungen der Versuch gemeinsam, Aufschlüsse über die der Lösung von Aufgaben zugrunde liegenden Denkprozesse zu gewinnen.

Formal wird dabei die Schwierigkeit von Aufgaben über das logistische Modell bestimmt. - Doch wenn man so festgestellt

hat, dass eine bestimmte Aufgabe z.B. 1,7 mal so schwer ist wie eine andere, wird man sich fragen müssen, warum die empirischen Parameter gerade so und nicht anders ausgefallen sind.

"Wenn es nun darum geht, Testmaterialien zu analysieren und dabei Aufschlüsse über die der Lösung zugrundliegenden Denkprozesse zu gewinnen, muss man versuchen, die Schwierigkeit oder psychologische Komplexität, welche im logistischen Modell durch einen Parameter σ gemessen wird, als Funktion von Struktureigenschaften der Aufgabe zu erklären." (FISCHER, 1974, S.361)

Wenn es einem gelingt, dafür eine befriedigende psychologische Theorie zu entwickeln, könnte man aufgrund der formalen Analyse von Denkaufgaben ihre psychologische Komplexität erkennen und vorhersagen. Damit könnte man auf der einen Seite über die reine Schwierigkeitsmessung heraus auch die Schwierigkeit erklären und sie auf ein allgemeines Prinzip zurückführen; andererseits könnten solche Analysen aber auch dazu benutzt werden, ein System von Regeln zu entwickeln, die die Konstruktion von Testaufgaben zur Messung einer umschriebenen Fähigkeit gestatten. Prinzipiell könnten sie dann auch zur automatischen Item konstruktion benutzt werden.

Eine der ersten Untersuchungen, nach einer Art "Baukastenprinzip" einen Test mit Items von vorhersagbarer Schwierigkeit bzw. Lösungswahrscheinlichkeit herzustellen, wurde von SCHEIBLECHNER (1972) durchgeführt.

Der Autor suchte dazu ein Testmaterial aus, bei dem die Aufgaben eine bestimmbare Anzahl von Anwendungen logischer Regeln erfordern. Dazu konstruierte er einen "Logik-Test" mit Aufgaben, die den Theoremen und Operationen des Aussagenkalküls entsprechen.

Um nun aber nicht zu überprüfen, ob den Probanden das Kalkül bekannt ist, wurden die Aufgaben nicht in der üblichen Schreibweise, sondern in einer entsprechenden graphischen Form vorgegeben. So ergibt sich z.B. für die Transformationsregel "nicht (p ∨ q)":

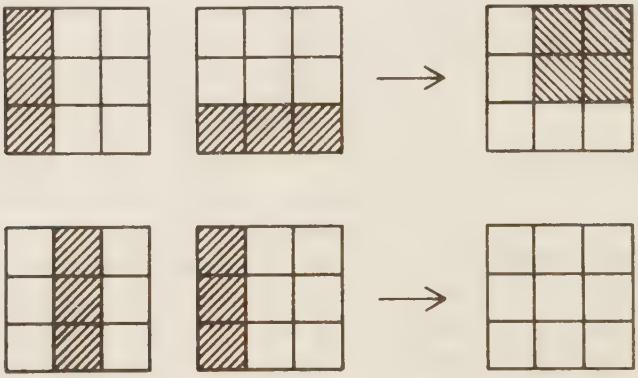


Abb. 21: Beispiel für eine Testaufgabe aus dem "Logik-Test" von SCHEIBLECHNER (1972). Im oberen Teil wird jeweils die Vorlage zur Erklärung der Regel (hier: "nicht (p ∨ q)" dargeboten, die im unteren Teil vom Probanden anzuwenden ist.

SCHEIBLECHNERs Modell zur Vorhersage der Aufgabenschwierigkeitsparameter enthält drei formale Operationen (Negation, Disjunktion und Konjunktion), einen inhaltlichen Parameter (Reihenfolge der Operationen) sowie drei Gedächtnisparameter (Zwischenresultat, Vertauschung, Kompliziertheit der Figur). Die Struktur dieses Modells lässt sich durch folgende Formel darstellen:

$$\sigma_i \approx t_i = \sum_{j=1}^{m} x_{i,j} \quad \delta_j + c \qquad \begin{array}{l} \text{für } i = 1,2,\ldots,k \\ \text{für } j = 1,2,\ldots,m \end{array}$$

Dabei bedeuten:

σ_i : logarithmisierter Aufgabenschwierigkeitsparameter der i-ten Aufgabe eines Tests;

t_i : theoretischer Aufgabenschwierigkeitsparameter;

$x_{i,j}$: Häufigkeit der j-ten elementaren Operation des Modells;

δ_j : Schwierigkeit der j-ten elementaren Operation;

d : Normierungskonstante.

Nach der Untersuchungsdurchführung mit 470 Schülern wurden für die einzelnen Items die Schwierigkeitsparameter geschätzt, und dann die multiple Regression mit dem oben angeführten Modell berechnet. Dabei ergab sich zwischen theoretischer und empirischer Schwierigkeit einer Korrelation von .92, die sich nach zusätzlicher Einsetzung von Übungsparametern sogar auf .9991 erhöhte.

In der Art dieses Ansatzes sind vor allem am Wiener Psychologischen Institut eine Reihe von Untersuchungen durchgeführt worden. Dabei analysierte FORMANN (1973) Aufgaben vom Typ der "Progressiven Matrizen" (RAVEN, 1956); HAUDEK (1972) arbeitete mit Aufgaben des "Zahlenreihenfortsetzens": SPADA, FISCHER & HEYNER (1974) verwandten Items aus der Mechanik und FISCHER (1973) untersuchte Mathematikaufgaben aus dem Gebiet der "Differentiation expliziter Funktionen".

In allen Untersuchungen konnte festgestellt werden, dass die Schwierigkeit der Testaufgaben mit einer für praktische Zwecke ausreichender Genauigkeit vorhergesagt werden konnte.

II. COMPUTER-UNTERSTÜTZTES TESTEN

1. EIN NEUER ANSATZ

Überblicken wir das Gesamtgebiet der computer-unterstützten Diagnostik, so können wir mit BÜRLI (1975) feststellen, dass die computer-unterstützte Testdurchführung im Gegensatz zur computer-unterstützten Testauswertung und Interpretation in den ersten Anfängen steht. Wie wir für den Bereich der die herkömmliche Testdurchführung "simulierende" automatisierte Testdurchführung festgestellt haben, können durch die Einbeziehung zusätzlicher Informationen - wie z.B. der zur Beantwortung der Testaufgaben benötigten Zeiten - Verbesserungen der diagnostischen Instrumente erzielt werden.

Doch müssen wir uns hier die Frage stellen, ob unter Abwägung des mit der Automatisierung herkömmlicher Testverfahren verbundenen Arbeitsaufwandes - von den finanziellen Aufwendungen einmal abgesehen - ein weiteres Fortschreiten in dieser Richtung überhaupt sinnvoll ist, d.h. ob versucht werden soll, durch die Einbeziehung weiterer Informationen die Messgenauigkeit und wahrscheinlich damit verbunden die Validität herkömmlicher Testverfahren schrittweise zu steigern, oder ob es nicht vielmehr sinnvoll ist, in diesem Anfangsstadium der Automatisierung von vornherein einen neuen Weg einzuschlagen.

Gegen eine - zwar mit diagnostischem Gewinn verbundene - allgemeine Automatisierung herkömmlicher Testverfahren sprechen die Schwächen der klassischen Testtheorie (s.z.B.

LORD & NOVICK, 1968; FISCHER, 1974) und die sich daraus ergebenden Schwächen der nach ihren Prinzipien konstruierten Testverfahren. Auch wenn hier zunächst noch wesentliche Fortschritte erzielt werden können, führt dieser Weg m.E. dennoch über kurz oder lang in eine Sackgasse, denn wir können zwar Verbesserungen im Detail erzielen, aber nicht damit die Schwächen der theoretischen Grundlagen kompensieren.

Für einen neuen Weg bei der computer-unterstützten Testdurchführung sprechen die Vorteile logistischer Messmodelle (s. wiederum LORD & NOVICK oder FISCHER) und speziell die Möglichkeit, mit Hilfe logistischer Messmodelle über den Ansatz des "massgeschneiderten Testens" zu einer präziseren und ökonomischen (vgl. GREEN, 1970) Bestimmung von Fähigkeitsparametern zu gelangen.

FISCHER (1974) führt als Nachteile des "tailored testing" auf der einen Seite die Notwendigkeit computer-gesteuerter-Itemauswahl und -Vorgabe, vor allem aber die Verfügbarkeit einer ausreichenden Menge homogener Testaufgaben an, und misst dieser Vorgehensweise wegen des aus diesen beiden Gründen erforderlichen Aufwandes eine relativ geringe praktische Bedeutung bei.

Mit dem von FISCHER angeführten Nachteil der computer-gesteuerten Itemauswahl und -Vorgabe brauchen wir uns an dieser Stelle nicht weiter zu beschäftigen, denn wir haben ja im experimentellen Teil dieser Arbeit u.a. aufzeigen können, dass die Entwicklung automatisierter Testsysteme auch unter einschränkenden technischen Bedingungen möglich ist.

Die Notwendigkeit, für das "tailored testing" eine grosse Menge von Items unterschiedlichster Schwierigkeitsgrade zur Verfügung haben zu müssen, bedeutet in der bisher üblichen Sichtweise von Testkonstruktion sicher einen wesentlichen Nachteil; denn es setzt einen enormen Arbeitsaufwand voraus, die Schwierigkeit von Testaufgaben jeweils zunächst in einer entsprechenden Probanden-Stichprobe empirisch bestimmen zu müssen. Wie wir jedoch bei der Diskussion der Aufgabenkonstruktion nach dem logistischen Modell zeigen konnten, ist es zumindest bei der Messung genau umschreibbarer Fähigkeiten (vgl. z.B. HAUDEK, 1972; FORMANN, 1973) möglich, die Schwierigkeit von Testaufgaben nicht "populationsabhängig" zu schätzen, sondern diese aufgrund ihrer Aufgabenstruktur zu bestimmen.

Eine solche formale Analyse der Schwierigkeit von Testaufgaben ermöglicht es aber auch, daraus ein System von Regeln zur Konstruktion von Aufgaben mit festgelegter Schwierigkeit zu entwickeln. Diese Regeln können nun ohne weiteres dazu benutzt werden, die Aufgabe der Itemkonstruktion einem Computer zu übertragen.

Für das Vorgehen beim "tailored testing" bedeutet das: Es ist nicht notwendig, eine grosse Menge von Items im Computer zu speichern, um sie bei Bedarf abzurufen, sondern diese können zu dem Zeitpunkt der Testdurchführung konstruiert werden, zu dem sie benötigt werden. - Dabei bleibt es unverständlich, warum diese Möglichkeiten z.B. von FISCHER (1974) noch nicht gesehen worden sind.

Bezogen auf die oben aufgeworfene Frage, welcher Weg zur Automatisierung der Testdurchführung im derzeitigen An-

fangsstadium der Automatisierung einzuschlagen ist, können wir also feststellen:

Wenn wir damit beginnen, die Testdurchführung zu automatisieren, so würde es zwar einen gewissen Fortschritt bedeuten, sich auf die Automatisierung herkömmlicher Testverfahren zu konzentrieren; grundsätzlich bessere Möglichkeiten werden aber erst dann eröffnet, wenn wir uns von dem klassischen Konzept lösen, und den Ansatz des "tailored-testing" aufgreifen.

2. DIE BEDEUTUNG DES NEUEN ANSATZES FÜR DIE DIAGNOSTIK

Zunächst mag die Vorstellung von einem Automaten, der parallel zur Testdurchführung die Konstruktion der vorzugebenden Testaufgaben übernimmt, als Wunschbild für eine weite Zukunft erscheinen, wenn man z.B. nur an den damit verbundenen Programmieraufwand denkt. Gehen wir jedoch zur Abschätzung der Realisierungsmöglichkeiten eines solchen Automaten auf die zu erarbeitenden Einzelschritte ein, so ergibt sich:

Wie in dieser Arbeit gezeigt wurde, ist die Entwicklung eines Systems zur automatisierten Testdurchführung auch unter beschränkenden technischen Gegebenheiten innerhalb absehbarer Zeiträume möglich.

Die theoretischen Grundlagen des "massgeschneiderten Testens liegen vor und können leicht in entsprechende Steuerungsprogramme übertragen werden. Ob sie auf die Dauer ausreichend sind, ist in der jetzigen Phase keine theoretische Frage sondern eine der empirischen Überprüfung.

Für eine Reihe genau definierter Fähigkeiten liegen die Konzepte einer aufgabenorientierten Schwierigkeitsbestimmung bereits vor. Damit lassen sich wenigstens diese Konzepte leicht in zumindest heuristische Aufgabenkonstruktionsprogramme übertragen; denn die Schwierigkeit der Programmerstellung liegt nicht in der Übertragung definierter Einzelschritte in eine computer-verständliche Sprache, sondern in der hinreichenden Analyse der Bedingungen.

Das bedeutet für die diagnostische Praxis, dass wir zumindest für solche klar definierten Fähigkeiten, wie sie etwa in der faktorenanalytischen Betrachtungsweise durch die THURSTONEschen Begriffe "Reasoning","Number" oder "Space"[1] beschrieben werden, zur Zeit in der Lage sind, automatisierte Testsystems nach dem hier beschriebenen Ansatz zu erstellen und sie als ökonomischere und präzisere Messinstrumente einzusetzen (vgl. LINN et al. 1969; CLEARY et al. 1968, 1968a, GREEN, 1970).

Der Diagnostiker wird damit von Routineaufgaben entlastet, die ein Automat besser und genauer erfüllen kann. Für die Praxis bedeutet dies eine verstärkte Zuwendung zum Gebiet der Beratung und Behandlung.

In der Forschung darf es nicht mehr darum gehen, zu jedem eng umschriebenen Problem und jede spezifische Population unter grossem Arbeitsaufwand ein spezielles Testverfahren zu erstellen. Hier ist vielmehr die Grundlagenforschung in dem Sinne weiterzuführen, dass zu den als grundlegend erkannten Fähigkeiten auch Konstruktionsprinzipien für die sie prüfenden Aufgaben abgeleitet werden können.

---

1) Die Ausgabe graphischer Darstellungen durch den Computer bis zur Erzeugung beweglicher Bilder (vgl. z.B. UTTAL, 1969) wird durch käuflich zu erwerbende Hilfsroutinen wesentlich erleichtert.

3. OFFENE FRAGEN

Über die "Reliabilität" von Testverfahren, die nach dem Konzept des "tailored testing" durchgeführt werden, ist zur Zeit noch wenig bekannt. Während in der klassischen Testtheorie über die Fehlervarianz eines Tests, der in einer bestimmten Population durchgeführt worden ist, die Messgenauigkeit des einzelnen Testergebnisses abgeschätzt wird, ermöglicht das logistische Modell die Abschätzung des Konfidenzintervalls für den Fähigkeitsparameter (FISCHER, 1974); praktische Anwendungen auf das "massgeschneiderte Testen" liegen ausser bei LORD (1971) bisher nicht vor. Hier sind die Grundlegungen ähnlicher Konzepte wie der klassischen "Reliabilität" bzw. "Validität" zu erarbeiten.

Neben der weiteren Aufbereitung der logistischen Messmodelle für die Praxis bedarf vor allem der gesamte diagnostische Prozess einer weiteren theoretischen Klärung. Erste Ansätze sind mit dem Modell von WESTMEYER (1972) gegeben; doch sind die dort vermittelten Anregungen bisher kaum aufgegriffen worden.

ZUSAMMENFASSUNG

Ausgehend von einer Reihe typischer Beispiele haben wir zunächst versucht, einen gewissen Überblick über die Einsatz- und Leistungsmöglichkeiten des Computers in verschiedenen Bereichen der Psychologie zu geben. Lässt man das grosse Gebiet der Datenanalyse unberücksichtigt, so lässt sich der Computer zunächst als genauer und standardisierter Experimentator einsetzen.

Neben der Textausgabe auf verschiedene Medien eröffnet vorallem die Bildschirmausgabe graphischer Darstellungen einschliesslich der Ausgabe bewegter Bilder (z.B. UTTAL, 1969) eine Vielzahl von Untersuchungsmöglichkeiten. Auch reichen die Möglichkeiten zur Registrierung von Versuchspersonenreaktionen bis hin zur genauen Zeitmessung. Weiter können die im Experiment anfallenden Daten mit einer sehr geringen Fehlerquote auf eine Reihe von Datenträgern ausgegeben, bzw. gleich nach Versuchsende verrechnet werden.

Über die reine Reizdarbietung und Reaktionsregistrierung hinaus werden die Möglichkeiten des Computers noch effektiver genutzt, wenn parallel zur eigentlichen Versuchsdurchführung die anfallenden Daten analysiert und aufgrund dieser Analyse der Versuchsfortgang gesteuert wird. Beispiele

finden sich im Bereich der Wahrnehmungspsychologie (KELLEY, 1969) und auf dem Gebiet der computer-unterstützten Unterweisung (STOLUROW, 1967; WEXLER, 1970).

Im engeren Bereich der psychologischen Diagnostik liegen für alle vier Teilbereiche des diagnostischen Prozesses (Durchführung, Auswertung, Interpretation und Beratung) auf der praktischen Ebene mehr oder weniger weit fortgeschrittene Automatisierungsversuche vor, für die jedoch zum Teil die theoretischen Grundlagen fehlen.

Bei der Testdurchführung beschränkt man sich bisher auf die Übernahme bereits bestehender Verfahren (z.B. ELWOOD, 1973) und die Untersuchung der affektiven Auswirkungen dieser Durchführungsart (z.B. LUSHENE, O'NEIL & DUNN, 1972).

Die numerische Testauswertung ist bisher am stärksten automatisiert (vgl. BÜRLI, 1973) und bedarf aufgrund der nur nach den Prinzipien der klassischen Testtheorie konzipierten Tests keiner eigenen Theorie.

Bei der qualitativen Testinterpretation liegen eine Reihe praktischer Ansätze vor (z.B. FOWLER, 1969, SLETTEN et al., 1972); die theoretischen Grundlegungen gehen entweder von der aktuarischen Betrachtungsweise des diagnostischen Urteilsprozesses (z.B. GOLDBERG, 1970), der klinischen Analyse (KLEINMUNTZ, 1970) oder von dem normativen Aspekt (WESTMEYER, 1972) aus.

Die Systeme zur automatisierten Beratung vermitteln zur Zeit fast nur Informationen (vgl. HARRIS, 1974); Versuche,

eine Theorie der Beratung als Grundlage für Beratungsprogramme einzusetzen, haben bisher nur modellhafte Bedeutung (ELLIS & TIEDEMANN, 1970).

Diese Arbeit beschäftigt sich im Rahmen der computer-unterstützten Testdiagnostik (CAT) mit dem engeren Gebiet der automatisierten Testdurchführung. Dabei soll in zwei Bereichen beispielhaft die Überlegenheit einer automatisierten Testdurchführung gegenüber der herkömmlichen nachgewiesen werden: Zum einen ist bei der Übertragung herkömmlicher Testverfahren die Möglichkeit zusätzlicher Informationsgewinnung (GREEN, 1970) am Beispiel der Latenzzeitmessung und der Einbeziehung der subjektiven Antwortsicherheit der Probanden (BOLDT, 1974) aufzuzeigen; zum anderen soll nachgewiesen werden, dass die Effektivität diagnostischen Vorgehens durch den Ansatz des "tailored testing" (LORD, 1970, 1971) wesentlich gesteigert werden kann.

Untersuchungen mit Systemen zur automatisierten Testdurchführung werden wesentlich dadurch erschwert, dass zwar die Prinzipien bisher erarbeiteter Lösungswege übernommen werden können, jedoch die technischen Voraussetzungen, die die einzelnen Untersucher vorfinden, so unterschiedlich sind, dass im Grunde der gesamte Lösungsweg jeweils neu zu erarbeiten ist.

Unter Berücksichtigung der für diese Untersuchung gegebenen technischen Möglichkeiten war deshalb zunächst ein allgemeines System zu entwickeln, das den Dialog zwischen Benutzer und Computer über Fernsehschirm und Eingabetastatur gestattet. Doch ergibt sich an dieser Stelle dann die psychologisch bedeutsame Frage, wie ein bestehendes Testverfahren in eine entsprechende automatisierte Form zu übertragen ist.

Während die bisherigen Untersuchungen zur automatisierten Testdurchführung implizit davon ausgehen, dass die dort gewählte Form der Übertragung die jeweils angemessene ist, (z.B. DUNN, LUSHENE & O'NEIL, 1972), sind wir von der inhaltlichen Analyse der Situation bei der herkömmlichen Durchführung eines "Papier-und-Bleistift"- Tests ausgegangen. Berücksichtigt man die aufgrund der technischen Voraussetzungen bestehenden Beschränkungen, so ergeben sich bei der herkömmlichen Darbietung zwei wesentliche Variable, die für die Vergleichbarkeit beider Durchführungsformen bedeutsam sein können: Die ständige Kontrolle über die bisher geleistete und noch zu leistende Arbeit und die Möglichkeit, während der Untersuchung einzelne Items zunächst zurückzustellen und sie später zu bearbeiten.

Aufgrund der experimentellen Überprüfung dieser beiden Variablen in der 1. Versuchsserie konnten wir feststellen, dass ohne ihre Berücksichtigung die Übereinstimmung zwischen beiden Durchführungsarten nicht zufriedenstellend ist. Ob nur einer oder beide Faktoren für eine hinreichende Übereinstimmung notwendig sind, konnte nicht genau entschieden werden.

Weiter ergab der Vergleich der Einschätzungen der beiden Versuchssituationen durch den Probanden, dass keine von beiden als stärker belastend erlebt wird. Die Vermutung, dass der einzelne Proband in der von ihm bevorzugten Versuchssituation die besseren Ergebnisse erzielt, konnte nicht bestätigt werden.

Mit der Entwicklung eines Systems zur automatisierten Testdurchführung und dem Nachweis der hinreichenden Übereinstimmung der Ergebnisse aus beiden Durchführungsarten, wa-

ren die Voraussetzungen zur Überprüfung des zusätzlichen diagnostischen Gewinnes aufgrund der bei der automatisierten Testdurchführung zusätzlich anfallenden Informationen geschaffen. Aufgrund der experimentellen Überprüfung (2. Versuchsserie) der beiden zusätzlichen Informationsmöglichkeiten der differenzierten Aufgabenbeantwortung und der Bestimmung der Latenzzeiten konnten wir anhand der Parameter Schwierigkeit, Trennschärfe und innere Konsistenz nachweisen, dass die automatisierte Durchführung herkömmlicher Verfahren abgesehen von einem möglichen ökonomischen Gewinn durch die wesentliche Erhöhung der Messgenauigkeit einen zusätzlichen diagnostischen Gewinn mit sich bringt. Dabei ist vor allem die Beachtung von Latenzzeiten von Bedeutung. Zwar hat diese für das einzelne Item bei weitem nicht die Bedeutung, die ihr oft zugeschrieben wird, doch trägt sie beträchtlich zur Erhöhung der Konsistenz des Gesamtverfahrens bei. Für die Praxis bedeutet das, dass bei der Automatisierung eines herkömmlichen Verfahrens auf jeden Fall die zusätzliche Informationsquelle der Latenzzeitmessung aufzugreifen ist, um durch eine nach Beantwortungszeiten differenzierte Aufgabenbewertung zu reliableren Gesamttestergebnissen zu kommen.

Wie wir beispielhaft aufzeigen konnten, lässt sich die Messgenauigkeit herkömmlicher, nach den Regeln der klassischen Testtheorie konstruierter Verfahren, durch die automatisierte Durchführung wesentlich steigern, womit wahrscheinlich auch eine Verbesserung ihrer Validität erzielt werden kann. Doch wenn man den mit der Automatisierung verbundenen Aufwand bedenkt, ist zu fragen, ob an dieser Stelle nicht grundsätzlich ein neuer Weg beschritten werden sollte.

Als der klassischen Testtheorie überlegene theoretische Grundlagen haben wir die logistischen Messmodelle eingeführt,

die vor allem in der Lernziel-orientierten Leistungsmessung immer stärker berücksichtigt werden, und auf die Vorteile des darauf aufbauenden Ansatzes des "massgeschneiderten Testens" (LORD, 1971), bei dem jeder Proband möglichst nur mit Testaufgaben geprüft wird, die seinen Fähigkeiten entsprechen, hingewiesen. Wie wir zeigen konnten, lässt sich das grösste Hindernis für die Einführung des "tailored testing", die Notwendigkeit eines grossen Pools homogener Testaufgaben, überwinden.Denn zumindest für eine Reihe klar definierbarer Fähigkeiten lässt sich die Schwierigkeit von Testaufgaben auf der Grundlage logistischer Modelle formal bestimmen (z.B. HAUDEK, 1972; FORMANN, 1973), und muss nicht empirisch festgelegt werden. Damit ermöglicht es die formale Analyse von Aufgabenschwierigkeiten, Regeln für die Konstruktion von Testaufgaben mit vorgegebener Schwierigkeit einzuführen, die ihrerseits zur automatischen Itemkonstruktion eingesetzt werden können. Für das "massgeschneiderte Testen" bedeutet das, dass nicht eine grosse Anzahl von Aufgaben im Computer gespeichert werden muss, sondern dass diese parallel zur Testdurchführung erstellt werden können.

Zumindest für eine Reihe eng umschriebener Fähigkeiten sind wir damit in der Lage, die Vorteile des "massgeschneiderten Testens" schon heute in der Praxis wahrzunehmen; diese Vorteile ergeben sich aus der präziseren Fähigkeitsbestimmung und der grösseren Ökonomie des Verfahrens.

Wenn wir also damit beginnen die Testdurchführung zu automatisieren - und dieser Ansatz ist notwendig, um den Psychologen von Routinearbeiten zu entlasten und ihn für seine eigentlichen Aufgaben der Beratung und Behandlung einsetzen zu können - so sollten wir uns meines Erachtens nicht damit zufrieden geben, bei den kurzfristig zu erzielenden Vor-

teilen der Automatisierung herkömmlicher Testverfahren stehen zu bleiben, sondern von vornherein diesen neuen Ansatz aufgreifen.

LITERATURVERZEICHNIS

AMTHAUER, R. (1970): Intelligenz-Struktur-Test. Göttingen.

APTER, M. (1973): The computer modelling of behavior, in: APTER, M. & WESTBY, G. (eds.): The computer in psycholgy.

APTER, M. & WESTBY, G. (eds.) (1973): The computer in psychology. London.

BELL, F.O., HOFF, A.L. & HOYT, K.B. (1964): Answer sheets do make a difference. Personnel Psychol.,17,65-71.

BHUSHAN, V. & GINTHER, J.R. (1968): Discriminating between a good and a poor essay. Behav.Sci.,13,417-420.

BIRNBAUM, A. (1968): Some latent trait models and their use in inferring an examinee's ability, in: LORD, F.M. & NOVICK, M.R. (eds.): Statistical theories of mental test scores.

BOLDT, R.F. (1974): An approximately reproducing scoring scheme that aligns random response and omission.Educ. Psychol.Measment.,34,56-61.

BOBROW, D.G. (1964): Natural language input for a computer problem solving system. Massachusetts Intitute of Technology Report MAC - TR - I. Ref.: HARRISON (1973).

BÜRLI, A. (1972): Automatische MMPI-Systeme. Schweiz.Z. Psychol.,31,281-291.

BÜRLI, A. (1973): Computer-unterstützte Testdiagnostik. Schweiz.Z.Psychol.,32,26-45.

BÜRLI, A. (1975): Computer-unterstützte und verzweigte Testdurchführung. Schweiz.Z.Psychol.,34,115-128.

CAHOON, D.D., PETERSON, L.P. & WATSON, C.G. (1968): Relative effectiveness of programmed text and teaching machine as a function of measured interests. J. Appl.Psychol.52, 454-456.

CASEY, R.G. & NAGY, G. (1971): Advances in pattern recognition. Scientific American,224,4,56-71.

CLEARY, T.A., LINN, R.L. & ROCK, D.A. (1968): An exploratry study of programmed tests. Educ.Psychol.Measment.,28, 345-360.

CLEARY, T.A., LINN, R.L. & ROCK, D.A. (1968a): Reproduction of total test score through the use of sequential programmed tests. J.Educ.Measment.,5,183-187.

COLBY, K.M., WEBER, S. & HILF, F.D. (1971): Artificial Paranoia. Artificial Intelligence,2. Ref.: MUSER, SEIDENSTÜCKER & SEIDENSTÜCKER (1974).

CROWDER, N.A. (1960): Automatic tutoring by intrinsic programming, in: LUMSDAINE, A.A. & GLASER, R. (eds.): Teaching machines and programmed learning: a source book. Washington, D.C..

DUNN, T.G., LUSHENE, R.E. & O'NEIL, H.F. (1972): Complete automation of the MMPI and a study of its response latencies. J.Consult.Clin.Psychol.,39,381-387.

DUNLOP, E. (1966): Essentials of the automated MMPI. Glendale.

EATON, M.E., ALTMAN, H., SCHUFF, S. & SLETTEN, I. (1970): Missouri automated psychiatric history for relatives and other informants. Diseases of the Nervous System,31, 198-202.

EINHORN, H.J. (1970): The use of nonlinear, noncompensatory models in decision making. Psychol.Bull.,73,221-230.

EINHORN, H.J. (1971): Use of nonlinear, noncompensatory models as a function of task and amount of information. Organ.Behav.Human Performance,6,1-28.

ELLIS, A.B. & TIEDEMANN, D.V. (1970): Can a machine counsel? in: HOLTZMAN, W.H. (ed.): Computer-assisted instruction, testing, and guidance.

ELWOOD, D.L. (1969): Automation of psychological testing. American Psychologist,24,287-289.

ELWOOD, D.L. (1972a): The retest reliability and cost analyses of automated and face-to-face intelligence testing. Int.J.Man-Machine Studies,4,1-22.

ELWOOD, D.L. (1972b): Automated WAIS testing correlated with face-to-face WAIS testing: a validity study. Int.J.Man-Machine Studies,4,129-137.

ELWOOD, D.L. (1972c): Validity of an automated measure of intelligence in borderline retarded subjects. Amer.J. Ment.Def.,77,90-94.

ELWOOD, D.L. (1973): Reliability of automated intelligence testing using a three-month test-retest interval. Int. Rev.Appl.Psychol.,22,157-163.

ELWOOD, D.L. & GRIFFIN, H.R. (1972): Individual intelligence testing without the examiner: reliability of an automated method. J.Consult.Clin.Psychol.,38,9-14.

EVAN, W.M. & MILLER, J.R. (1969): Differntial effects on response bias of computer vs. conventional administration of a social science questionnaire. Behav.Sci.,14,216-227.

FELDMAN, J. (1959): Simulation of behavior in the binary choice experiment. Proc. Western Joint Computer Conference, 16,133-144.

FINNEY, J.C. (1966): A programmed interpretation of the MMPI and the CPI. Arch.Gen.Psychiatry,15,75-81.

FINNEY, J.C. (1967): Methodological problems in programmed composition of psychological test reports. Behav.Sci.,12, 142-152.

FISCHER, G.H. (1973): The linear logistic test model as an instrument in educational research. Acta Psychol.,37, 359-374.

FISCHER, G.H..(1974): Einführung in die Theorie psychologischer Tests. Bern.

FORMANN, A. (1973): Die Konstruktion eines neuen Matrizentest und die Untersuchung des Lösungsverhaltens mit Hilfe des linearen logistischen Testmodells. Unver. Diss. Psychol. Inst. Univ. Wien.

FOWLER, R.D. (1964): Computer processing and reporting of personality test data. Proc.Ann.Conv.Amer.Psychol.Ass..

FOWLER, R.D. (1967): Computer interpretation of personality test data: the automated psychologist. Compr.Psychiatry, 8,455-467

FOWLER, R.D. (1967a): MMPI computer interpretation for college counseling. Proc.Ann.Conv.Amer.Psychol.Ass.,363-364.

FOWLER, R.D. (1968): MMPI computer interpretation for college counseling. J.Psychol.,69,201-207.

FOWLER, R.D. (1969): The current status for computer interpretation of psychological tests. Amer.J.Psychiatry,125, 21-27.

FOWLER, R.D. & COYLE, F.A. (1968): Computer aplication of MMPI actuarial systems with a college population. J.Clin. Psychol.,69,233-236.

FOWLER, R.D. & MARLOWE, G.H. (1968a): A computer program for personality analysis. Behav.Sci.,13,413-416.

FOWLER, R.D. MILLER, M.L. (1969): Computer interpretation of the MMPI: Its use in clinical practice. Arch.Gen. Psychiatry,21,502-508.

FOWLER, R.D., STEVENS, S.S., COYLE, F.A. & MARLOWE, G.H. (1968b): Comparsion of two methods of identifying maladjusted students. J.Psychol.,69,165-168.

FRICKE, R. (1972): Über Messmodelle in der Schulleistungsdiagnostik. Düsseldorf.

FRIJDA, N.H. (1967): Problems of computer simulation. Behav. Sci.,12,59-67.

GEHRING, A. & GLASER, P. (1972): Ein Computerprogramm zur Auswertung und Interpretation des MMPI. Schweiz.Z.Psychol., 31,292-297.

GLAUBITZ, G. (1975): Möglichkeiten und Probleme computergesteuerten automatischen Testens. Unv.Zulassungsarbeit zur Diplom-Hauptprüfung für Psychologen. Psychol. Inst. Univ. Münster.

GLUECK, B.C. & REZNIKOFF, M. (1965): Comparison of computerderived personality profile and projective psychological test findings. Amer.J.Psychiatry,121,1156-1161.

GOLDBERG, L.R. (1965): Diagnosticans vs. diagnosis signs: The diagnosis of psychoses vs. neuroses from the MMPI. Psychol.Mon.,79 (Whole No. 602).

GOLDBERG, L.R. (1968): Simple models or simple processes? Some research of clinical judgement. Amer.Psychologist, 23,483-496.

GOLDBERG, L.R. (1970): Man vs. model of man: A rationale plus some evidence for a method of improving of clinical inferences. Psychol.Bull.,73,422-432.

GOLDBERG, J.B. (1966): Computer analysis of sentence completions. J.Proj.Techn.Pers.Assesm.,30,37-45.

GREEN, B.F. (1968): Descriptions and explanations: A comment on pap ers by Hoffman and Edwards, in: KLEINMUNTZ, B. (ed.): Formal representation of human judgement. New York.

GREEN, B.F. (1970): Comments on tailored testing, in: HOLTZMAN, W.H. (ed.): Computer-assisted intruction, testing, and guidance.

GRUBB, R.E. & SELFRIDGE, R.E. (1963): The computer tutoring of statistics. Computer and Automation,13,20-26.

GULLIKSEN, H. (1950): Theory of mental tests. New York.

HARRIS, J. (1974): The computer: Guidance tool of the future. J.Counsel.Psychol.,21,331-339.

HARRISON, G. (1973): The computer in the psychology of language, in: APTER & WESTBY (eds.): The computer in psychology.

HATHAWAY, S.R. & MC KINLEY, J.C. (1967): Minnesota multiphasic personality inventory: Manual for administration and scoring. New York.

HAUDEK, U.R. (1973): Ein Modell der Schwierigkeit von Denkaufgaben. Unv.Diss.Psychol.Inst.Univ. Wien.

HEDL, J.J., O'NEIL, H.F. & HANSEN, D.N. (1973): Affective reactions toward computer-based intelligence testing. J.Consult.Clin.Psychol.,40,217-222.

HILF, D.F., COLBY,K.M., WITTNER, W.K.,& HALL, A.W. (1971): Machine mediated interview. J.nerv.ment.Dis.,152,278-288.

HOFFMAN, P.J. (1960): The paramorphic representation of clinical judgement. Psychol.Bull.,57,116-131.

HOFFMAN, P.J. (1968): Cue-consistency and configurality in human judgement, in: KLEINMUNTZ, B. (ed.): Formal representation of human judgement. New York.

HOLTZMAN, W.H. (1970): Computers in education, in: HOLTZMAN, W.H. (ed.): Computer-assisted intruction, testing, and guidance. New York.

JOHNSON, D.F. & MIHAL, W.L. (1973): Performance of blacks and whites in computerized versus manual testing environments. American Psychologist,28,694-699.

JUDD, W. & GLASER, R. (1969): Response latency as a function of training method, information level, acquisition, and overlearning. J.Educ.Psychol., Monogr.,part 2,1-30.

KELLEY, C.R. (1969): Adaptive and automated research techniques from engineering psychology. American Psychologist, 24,293-297.

KLEINMUNTZ, B. (1963): MMPI decision rules for the indentification of college maladjustment: A digital computer approach. Psychol.Monogr.,77.

KLEINMUNTZ, B. (1963a): Personality test interpretation by digital computer. Science,139,416-418.

KLEINMUNTZ, B. (1963b): A portrait of the computer as a young clinican. Behav.Sci.,8,154-156.

KLEINMUNTZ, B. (1963c): Profile analysis revisited: A heuristic approach. J.Counsel.Psychol.,10,315-324.

KLEINMUNTZ, B. (1967): Sign or seer: Another example. J.abnormal Psychol.,72,163-165.

KLEINMUNTZ, B. (1968): The processing of clinical information by man and machine, in: KLEINMUNTZ, B. (ed.): Formal representation of human judgement. New York.

KLEINMUNTZ, B. (1969): Personality assessment ba computer. Science J.,59-64.

KLEINMUNTZ, B. (1969a): Clinical information processing by computer: an essay and selected readings. New York.

KLEINMUNTZ, B. (1970): Clinical information processing by computer, in: NEWCOMB et al. (eds.): New directions in psycholgy. New York.

KOSON, D., KITCHEN, C., KOCHEN, M. & STOLOSKY, D. (1970): Psychological testing by computer. Educ.Psychol.Measment., 30,803-810.

KRÜGER, K., HYLLA, E., BARGMANN, R. & HORN, W. (1965): Diagnostischer Rechentest "Zahlenrechnen ZR 4+" für 4. und höhere Klassen. Weinheim.

LANGHEINRICH, D. (1975): Computer im psychologischen Experiment, in: HAWEL, W. (ed.): Datenverarbeitung in der Psychologie. Stuttgart.

LAUGHERY, K.R. & GREGG, L.W. (1962): Simulation of human problem solving. Psychometrika,27,265-282.

LEVY, L.H. (1963): Psychological interpretation. New York.

LIENERT, G.A. (1969): Testaufbau und Testanalyse. Weinheim.

LINN, R.L., ROCK, D.A. & CLEARY, T.A. (1969): The development and evaluation of several programmed testing methods. Educ. Psychol.Measment.,29,129-146.

LORD, F.M. (1970): Some test theory for tailored testing, in: HOLTZMAN, W.H. (ed,): Computer-assisted instruction, testing, and guidance.

LORD, F.M. (1971): Robbins-Monro procedures for tailored testing. Educ.Psychol.Measment.,31,3.31.

LORD, F.M. & NOVICK, M.R. (1968): Statistical theories of mental test scores. Princeton.

LUSHENE, R.E., O'NEIL, H.H. & DUNN, T. (1972): Equivalent validity of a completely computerized MMPI. CAI Tech. Memo, Florida State Univ., Ref.: BÜRLI (1975).

MEEHL, P.E. (1954): Clinical versus statistical prediction. Minneapolis.

MOSER, U., SEIDENSTÜCKER, G. & SEIDENSTÜCKER, E. (1974): Interviewforschung: Modelle des klinischen Gesprächs (Therapie und Interview), in: SCHRAML, W.J. &BAUMANN, U. (eds.): Klinische Psychologie II. Bern.

MOSER, U., v. ZEPPELIN, I. & SCHNEIDER, W. (1970): Computer simulation of a model of neurotic defense processes. Behav.Sci.,15,194-202.

NAYLOR, J.C. & WHERRY, R.J. (1965): The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. Educ.Psychol.Measment.,25,969-986.

NEWELL, A., SHAW, J.C. & SIMON, H.A. (1959): A general problem solving program for a computer. Computers and Auomation,8,10-16.

NEWELL, A. & SIMON, H.A. (1961): GPS, a program that simulates human thought, in: BILLING, H. (ed.): Lernende Automaten. München.

PAGE, E.B. (1968): The use of the computer in analysing student essays. Review of Educ,,14,210-224.

PLATMAN, S.R. & WEINSTEIN, B. (1970): The diagnosis game. Dis.Nerv.Sys.,31,561-566.

PRICE, G.E. (1974): Counselor and computer effectiveness in helping students select courses. J. Counsel.Psychol.,21, 351-354.

PRICE, G.E. & GILMORE, R.G. (1973): Computer assisted course selection. J. Counsel.Psychol.,20,327-331.

RASCH, G. (1960): Probabilistic models for some intelligence and attainment tests. Kopenhagen.

RASCH, G. (1966): An item analysis which takes individual differences in account. Brit.J.Math.Stat.Psychol.,19,49-57.

RAVEN, J.C. (1956): Standard progressive matrices. London.

ROBINSON, J.O. (1973): The computer in clinical psychology, in: APTER & WESTBY (eds.): The computer in psychology.

SARBIN, T.R., TAFT, R. & BAILEY, D.E. (1960): Clinical inference and cognitive theory. New York.

SCHEIBLECHNER, H. (1972): Das Lernen und Lösen komlpexer Denkaufgaben. Z.exp.angew.Psychol.,19,476-506.

SCHULZ, D. (1975): Zur Fähigkeit des Diagnostikers linear bzw. non-linear zu kombinieren. Z.exp.engew.Psychol., 22,138-173.

SEIDENSTÜCKER, G. (1971): Funktionen der Computersimulierung in der psychologischen Theorienbildung. Regensburg (im Manuskript).

SHUFORD, E.H., ALBERT, A. & MASSENGILL, H.E. (1966): Admissible probabiltity measurement procedures. Psychometrika,31,125-145.

SLETTEN, I.W., ERNHART, C.B. & ULETT, G.A. (1970a): The Missouri automated mental status examination: Development, use, and reliability. Comprehen.Psychiatry,11,315-327.

SLETTEN, I.W., ULETT, G.A., ALTMAN, H. & SUNDLAND, D. (1970b): The Missouri standard system of psychiatry (SSUP): Computer generated diagnosis. Arch.Gen.Psychiatry,23,73-79.

SLETTEN, I.W., ALTMAN, H. & ULETT, G.A. (1971): Routine diagnosis by computer. Amer.J.Psychiatry,127,1147-1152.

SLETTEN, I.W., SCHUFF, S., ALTMAN, H. & ULETT, G.A. (1972): A stadewide computerized psychiatric system: Demographic, diagnostic and mental status data. Int.J.Soc.Psychiatry, 18,30-40.

SLOSSEN, R.L. (1963): The Slossen intelligence test for children and adults. East Aurora, N.Y.

SMITH, M.S. (1968): The computer and the TAT. J.School Psychol.,6,206-214.

SMITH, R.E. (1963): Examination by computer. Behav.Sci.,8, 76-79.

SPADA, H., FISCHER, G.H. & HEYNER, W. (1974): Denkoperationen und Lernprozesse bei der Lösung von Problemstellungen aus der Mechanik, in: ECKENSBERGER, L.H. (ed.): Bericht über den 28. Kongress der DGfP, Saarbrücken 1972. Göttingen.

SPIELBERGER, C.D. GORSUCH, R.L. & LUSHENE, R.E. (1970): Manual for the state-trait anxiety inventory. Palo Alto.

SPITZER, R.L. & ENDICOTT, J. (1968): A computer program for psychiatric diagnosis utilizing the differential diagnostic procedure. Arch.Gen.Psychiatry,18,746-756.

SPITZER, R.L. & ENDICOTT, J. (1969): Diagno II: Further development in a computer program for psychiatric diagnosis. Amer.J.Psychiatry,125,Jan.supp.,12-21.

STATISTICAL RESEARCH GROUP OF COLUMBIA UNIVERSITY (1945): Sequential analysis of statistical data. New York. Ref.: GREEN (1970).

STOLUROW, L.M. (1965): A model and cybernetic system for research on the teaching-learning process. Programmed Learning and Educational Technology,2,138-157.

STOLUROW, L.M. (1967): A computer assisted instructional system in theory and research, in: UNWIN, D. & LEEDHAM, J. (eds.): Aspects of educational technology. London.

STONE, P.J., DUNPHY, C.D., SMITH, M.S. & OGLIVIE, D.M. (1967): The general inquirer: A computer approach to content analysis in the behavioral sciences. Cambridge, Mass..

SUPER, D.E. (1970): Of Tiedeman, Ellis, and machines that become media when given the right message, in: HOLTZMAN, W.H. (ed.): Computer-assisted intruction, testing, and guidance.

SUPPES, P., JERMAN, M. & BRIAN, D. (1968): Computer assisted instruction: Stanford's 1965-66 arithmetic program. New York.

SUPPES, P. & MORNINGSTAR, M. (1970): Stanford programs in arithmetic, logic, and Russian, in: HOLTZMAN, W.H. (ed.): Computer-assisted instruction, testing, and guidance.

SWENSON, W.M., ROME, H.P., PEARSON, J.S. & BRANNICK, T.L. (1965): A totally automated psychological test. J.Amer. Med.Ass.,191,925-927.

TOBIAS, S., HEDL, J.J.,& TOWLE, N.J. (1974): Response time and test anxiety. Psychol.Rep.,34,479-485.

UTTAL, W.R. (1967): Real-time computers: Technique and applications in the psychological sciences. New York.

UTTAL, W.R. (1969): Buggywhips, whalebones, and clippboards: Some notes on generating complex stimuli with small computers. Amer.Psychologist,24,202-206.

VELDMAN, D.J. (1967): Computer-based sentence-completion interviews. J.Counsel.Psychol.,14,153-157.

VUKOVICH, A.F. (1964): Die Konstruktion psychologischer Tests, in: HEISS, R., GROFFMANN, K.J. & MICHEL, L. (eds.): Hdb. d. Psychol., Bd. 6, Psychologische Diagnostik. Göttingen.

WATERS, C.F. & BAYROFF, A.G. (1971): A comparsion of computer-simulated conventional and branching tests. Educ.Psychol. Measment.,31,125-136.

WEBB, J.T., MILLER, M.L. & FOWLER, R.D. (1970): Extending professional time: A computerized MMPI interpretation service. J.Clin.Psychol.,26,210-214.

WEBB, J.T. (1970): Validity and utility of computer-produced MMPI reports with veterans administration psychiatric population. Proc.Ann.Conv.Amer.Psychol.Ass.,5,541-542.

WEIZENBAUM, J. (1966): ELIZA - a computer program for the study of natural language communications. Comm. ACM,9, 36-45.

WESTMEYER, H. (1972): Logik der Diagnostik. Stuttgart.

WEXLER, J.D. (1970): A teaching program that generates simple arithmetic problems. Int. J. Man-Machine Stud.,2,1-27.

WIGGINS, N. (1971): Individual differnces in diagnostic judgements of psychosis and neurosis from the MMPI. Educ. Psychol.Measment.,31,199-214.

WIGGINS, N. & KOHEN, E. (1971): Man vs. model of man revisited: The forecasting of gradual school success. J.Pers. Soc. Psychol.,19,100-106.

WIDROW, B. & SMITH, F.W. (1964): Pattern-recognizing control-systems, in: TOU, J.T. & WILCOX, R.H. (eds.): Computer and information sciences: Collected papers on learning, adaptation and control in information systems. Washington.

TABELLENANHANG

Hier sind aus der 2. Versuchsserie folgende Tabellen zusammengestellt:

Versuchsgruppe (1)

Schwierigkeitsindices, Schwierigkeitsparameter, mittlere Latenzzeiten und Streuungen bei richtiger bzw. falscher Lösung der einzelnen Aufgaben sowie Korrelationen zwischen Lösung und dafür benötigter Zeit für die Untertests:

SATZERGANZUNG	Tab. A 1
WORTAUSWAHL	Tab. A 2
ANALOGIEN	Tab. A 3

Versuchsgruppe (2)

Mittelwerte und Streuungen der Punktwerte, Schwierigkeitsparameter, Mittelwerte und Streuungen der Latenzzeiten sowie Korrelationen zwischen Punktwerten und Latenzzeiten für die Items der Untertests:

SATZERGÄNZUNG	Tab. A 4
WORTAUSWAHL	Tab. A 5
ANALOGIEN	Tab. A 6

Häufigkeiten der Punktwerte, Summen und prozentualer Anteil für die 5 Antwortmöglichkeiten der Items für die Untertests:

SATZERGANZUNG	Tab. A 7
WORTAUSWAHL	Tab. A 8
ANALOGIEN	Tab. A 9

Item	Schwierigkeit P(i)	b(i)	Latenzzeit Gesamt-Mittel	Sigma	Mittel bei r	Mittel bei f	Korrelation
1	.75	-1.09	21.95	7.97	20.20	27.20	-0.35
2	.75	-1.09	23.80	10.89	23.26	25.40	-0.08
3	.80	-1.38	17.80	9.88	15.75	26.00	-0.41
4	.90	-2.19	29.25	15.88	29.50	27.00	0.04
5	.95	-2.94	23.20	10.92	23.26	22.00	0.02
6	.60	-0.40	26.20	17.52	27.08	24.87	0.06
7	.70	-0.84	20.90	13.32	14.42	36.00	-0.74***
8	.45	0.20	17.55	9.61	22.11	13.81	0.42
9	.60	-0.40	30.25	13.98	33.50	25.37	0.28
10	.90	-2.19	30.20	22.11	32.00	15.00	0.23
11	.70	-0.84	18.85	6.40	18.42	19.83	-0.10
12	.90	-2.19	34.75	23.99	33.72	44.00	-0.12
13	.70	-0.84	29.50	26.62	25.28	39.33	-0.24
14	.65	-0.61	26.60	15.25	21.76	35.57	-0.43*
15	.40	0.40	35.30	31.92	44.75	29.00	0.24
16	.50	0.0	28.45	20.87	26.40	30.50	-0.09
17	.45	0.20	25.70	11.68	26.44	25.09	0.05
18	.75	-1.09	23.90	22.25	27.40	13.40	0.27
19	.95	-2.94	13.80	7.29	14.00	10.00	0.11
20	.30	0.84	34.20	34.26	24.33	38.42	-0.18

Tab. A 1: Schwierigkeitsindices (P), Schwierigkeitsparameter (b), mittlere Latenzzeiten und Streuungen, mittlere Zeiten bei richtiger (r) bzw. falscher (f) Lösung und Korrelationen zwischen Lösung der Aufgabe und dafür benötigter Zeit für die Items des IST-Untertests SATZERGÄNZUNG in der Versuchsgruppe (1)

Item	Schwierigkeit P(i)	b(i)	Latenzzeit Gesamt-Mittel	Sigma	Mittel bei r	Mittel bei f	Korrelation
1	1.00	-	11.75	4.21	-	-	-
2	.95	-2.94	14.35	7.12	14.42	13.00	0.04
3	.35	0.61	33.80	19.79	31.74	34.92	-0.07
4	.90	-2.19	22.35	17.06	19.88	44.50	-0.43*
5	.80	-1.38	20.45	12.82	21.68	15.50	0.19
6	.90	-2.19	16.25	11.80	14.61	31.00	-0.41
7	.95	-2.94	32.35	29.89	30.26	72.00	-0.30
8	.90	-2.19	23.50	28.22	24.27	16.50	0.08
9	.75	-1.09	16.65	8.74	15.73	19.40	-0.18
10	.45	0.20	36.10	21.66	34.44	37.45	-0.06
11	.35	0.61	34.65	20.91	45.42	28.84	0.37
12	.70	-0.84	13.35	5.62	14.50	10.66	0.31
13	.50	0.0	23.95	15.81	18.60	29.30	-0.35
14	.45	0.20	24.30	9.71	25.11	23.63	0.07
15	.80	-1.38	24.15	20.53	27.62	10.25	0.33
16	.60	-0.40	36.15	32.36	23.41	55.25	-0.48*
17	.60	-0.40	22.40	18.10	24.50	19.25	0.14
18	.50	0.0	21.00	10.67	23.10	18.90	0.19
19	.15	1.73	38.05	34.54	33.33	38.88	-0.05
20	.30	0.84	27.10	20.41	40.83	21.21	0.44*

Tab. A 2: Schwierigkeitsindices (P), Schwierigkeitsparameter (b), mittlere Latenzzeiten und Streuungen, mittlere Zeiten bei richtiger (r) und falscher (f) Lösung und Korrelationen zwischen Lösung der Aufgabe und dafür benötigter Zeit für die Items des IST-Untertests WORTAUSWAHL in der Versuchsgruppe (1)

Item	Schwierigkeit P(i)	b(i)	Latenzzeit Gesamt-Mittel	Sigma	Mittel bei r	Mittel bei f	Korrelation
1	1.00	-	8.45	3.08	-	-	-
2	1.00	-	35.55	34.84	-	-	-
3	1.00	-	13.10	6.45	-	-	-
4	.90	-2.19	21.30	10.96	21.11	23.00	-0.05
5	.55	-0.20	38.75	29.70	35.00	43.33	-0.13
6	.95	-2.94	16,75	9.26	16.68	18.00	-0.03
7	.95	-2.94	18.05	8.33	18.42	11.00	0.19
8	.60	-0.40	26.80	13.87	24.75	29.87	-0.18
9	.70	-0.84	32.70	26.57	22.64	56.16	-0.57**
10	.85	-1.73	30.60	16.45	29.76	35.33	-0.12
11	.80	-1.38	30.40	27.19	29.43	34.25	-0.07
12	.85	-1.73	29.00	18.47	26.52	43.00	-0.31
13	.90	-2.19	21.40	16.72	21.61	19.50	0.03
14	.65	-0.61	38.85	17.73	31.76	52.00	-0.54
15	.75	-1.09	29.70	18.71	25.93	41.00	-0.34
16	.40	0.40	35.60	18.05	34.00	38.33	-0.11
17	.65	-0.61	63.00	51.85	65.38	58.57	0.06
18	.80	-1.38	32.85	17.28	32.56	34.00	-0.03
19	.40	0.40	44.05	21.81	43.50	44.41	-0.02
20	.25	1.09	32.75	14.48	33.80	32.40	0.04

Tab. A 3: Schwierigkeitsindices (P), Schwierigkeitsparameter (b), mittlere Latenzzeiten und Streuungen, mittlere Zeiten bei richtiger (r) und falscher (f) Lösung und Korrelationen zwischen Lösung der Aufgabe und dafür benötigter Zeit für die Items des IST-Untertests ANALOGIEN in der Versuchsgruppe (1)

Item	Punktwert Mittel	Sigma	b(i)	Latenzzeit Mittel	Sigma	Korrelation
1	2.40	0.94	0.08	65.80	16.89	0.09
2	1.65	1.18	0.70	59.60	29.15	-0.14
3	4.15	1.53	-1.58	41.40	19.51	-0.05
4	4.20	1.32	-1.66	58.50	21.85	-0.22
5	3.65	1.78	-1.00	59.55	35.61	-0.31
6	2.80	1.64	-0.24	50.70	21.37	-0.23
7	4.30	1.71	-1.82	31.05	13.60	-0.71***
8	2.10	2.33	0.32	39.50	17.61	-0.08
9	2.00	2.17	0.40	62.85	42.71	-0.14
10	4.00	1.68	-1.38	46.05	21.14	-0.18
11	3.90	1.71	-1.26	34.35	10.38	-0.07
12	2.35	2.08	0.12	57.55	31.96	-0.46*
13	2.40	1.95	0.08	37.40	15.01	-0.26
14	1.90	1.74	0.48	48.25	19.10	-0.15
15	1.75	1.99	0.62	52.60	26.17	-0.13
16	2.55	1.76	-0.04	38.15	10.99	-0.29
17	2.35	2.00	0.12	42.20	18.44	-0.28
18	2.45	2.03	0.04	35.05	14.24	-0.25
19	3.40	1.50	-0.76	35.95	14.16	-0.16
20	0.80	1.36	1.66	38.15	15.34	0.12

Tab. A 4: Mittelwerte und Streuungen der Punktwerte, Schwierigkeitsparameter (b), Mittelwerte und Streuungen der Latenzzeiten und Korrelationen zwischen Punktwerten und Latenzzeiten für die Items des IST-Untertests SATZERGÄNZUNG in der Versuchsgruppe (2)

Item	Punktwert Mittel	Sigma	b(i)	Latenzzeit Mittel	Sigma	Korrelation
1	4.35	1.59	-1.90	34.15	25.24	-0.95***
2	4.75	0.78	-2.94	38.20	19.89	-0.44*
3	2.20	2.41	0.24	69.00	68.03	-0.35
4	4.25	1.83	-1.74	32.50	18.61	-0.03
5	4.00	2.05	-1.38	24.95	14.14	-0.53*
6	4.85	0.67	-3.48	21.60	8.28	-0.13
7	4.50	1.53	-2.20	42.45	29.79	0.21
8	3.65	2.20	-1.00	30.95	16.94	-0.23
9	3.85	2.08	-1.20	26.80	17.39	-0.33
10	2.50	2.37	0.0	50.45	41.08	-0.02
11	2.15	2.43	0.28	41.95	27.63	0.04
12	2.00	2.51	0.40	29.70	11.36	-0.11
13	2.75	2.55	-0.20	36.60	20.24	-0.18
14	2.55	2.52	-0.04	36.75	28.03	0.09
15	3.85	1.98	-1.20	47.50	33.04	-0.16
16	3.80	2.14	-1.16	55.85	60.80	-0.53*
17	1.50	2.35	1.38	39.25	31.47	-0.33
18	3.10	2.42	-0.48	51.05	38.95	0.21
19	2.35	2.47	0.12	68.20	54.01	-0.31
20	1.20	2.06	1.82	40.05	32.40	0.10

Tab. A 5: Mittelwerte und Streuungen der Punktwerte, Schwierigkeitsparameter (b), Mittelwerte und Streuungen der Latenzzeiten und Korrelationen zwischen Punktwerten und Latenzzeiten für die Items des IST-Untertests WORTAUSWAHL in der Versuchsgruppe (2)

Item	Punktwert Mittel	Sigma	b(i)	Latenzzeit Mittel	Sigma	Korrelation
1	5.00	0.0	-	19.80	6.77	-
2	4.25	1.83	-1.74	45.20	46.10	-0.08
3	4.60	1.27	-2.44	25.05	14.75	-0.19
4	4.25	1.83	-1.74	34.35	23.40	-0.37
5	4.40	1.31	-2.00	51.60	30.86	-0.29
6	4.65	1.18	-2.58	24.80	12.84	-0.14
7	4.50	1.53	-2.20	28.70	13.69	-0.13
8	2.85	2.34	-0.28	43.40	23.41	-0.52*
9	3.90	1.99	-1.26	40.10	19.76	-0.66**
10	3.20	2.11	-0.58	40.75	19.79	-0.14
11	3.80	2.06	-1.16	33.35	15.17	-0.02
12	3.60	2.23	-0.94	57.50	57.10	0.10
13	4.05	1.95	-1.44	37.50	41.08	-0.02
14	3.30	2.29	-0.66	56.95	27.18	-0.03
15	3.25	2.26	-0.62	41.25	25.80	-0.33
16	1.75	2.44	0.62	41.65	18.45	-0.42
17	2.50	2.56	0.0	53.35	27.90	0.02
18	4.00	2.05	-1.38	57.70	33.37	-0.42
19	1.55	2.32	0.80	59.90	25.09	-0.15
20	1.45	2.18	0.90	50.75	23.12	-0.49*

Tab. A 6: Mittelwerte und Streuungen der Punktwerte, Schwierigkeitsparameter (b), Mittelwerte und Streuungen der Latenzzeiten und Korrelationen zwischen Punktwerten und Latenzzeiten für die Items des IST-Untertests ANALOGIEN in der Versuchsgruppe (2)

ITEM 1

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	3	14	2	0	0	1	23	23.0
2 (F)	I	4	9	7	0	0	0	23	23.0
3 (F)	I	16	4	0	0	0	0	4	4.0
4 (F)	I	18	2	0	0	0	0	2	2.0
5.(R)	I	1	1	9	7	2	0	8	8.0

ITEM 2

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	6	7	5	1	0	1	25	25.0
2 (F)	I	10	6	3	0	0	1	17	17.0
3 (F)	I	6	13	1	0	0	0	15	15.0
4.(R)	I	3	6	8	2	0	1	33	33.0
5 (F)	I	15	5	0	0	0	0	5	5.0

ITEM 3

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	1	0	0	0	1	6	6.0
2.(R)	I	1	1	1	2	1	14	83	83.0
3 (F)	I	19	1	0	0	0	0	1	1.0
4 (F)	I	15	2	2	1	0	0	9	9.0
5 (F)	I	19	1	0	0	0	0	1	1.0

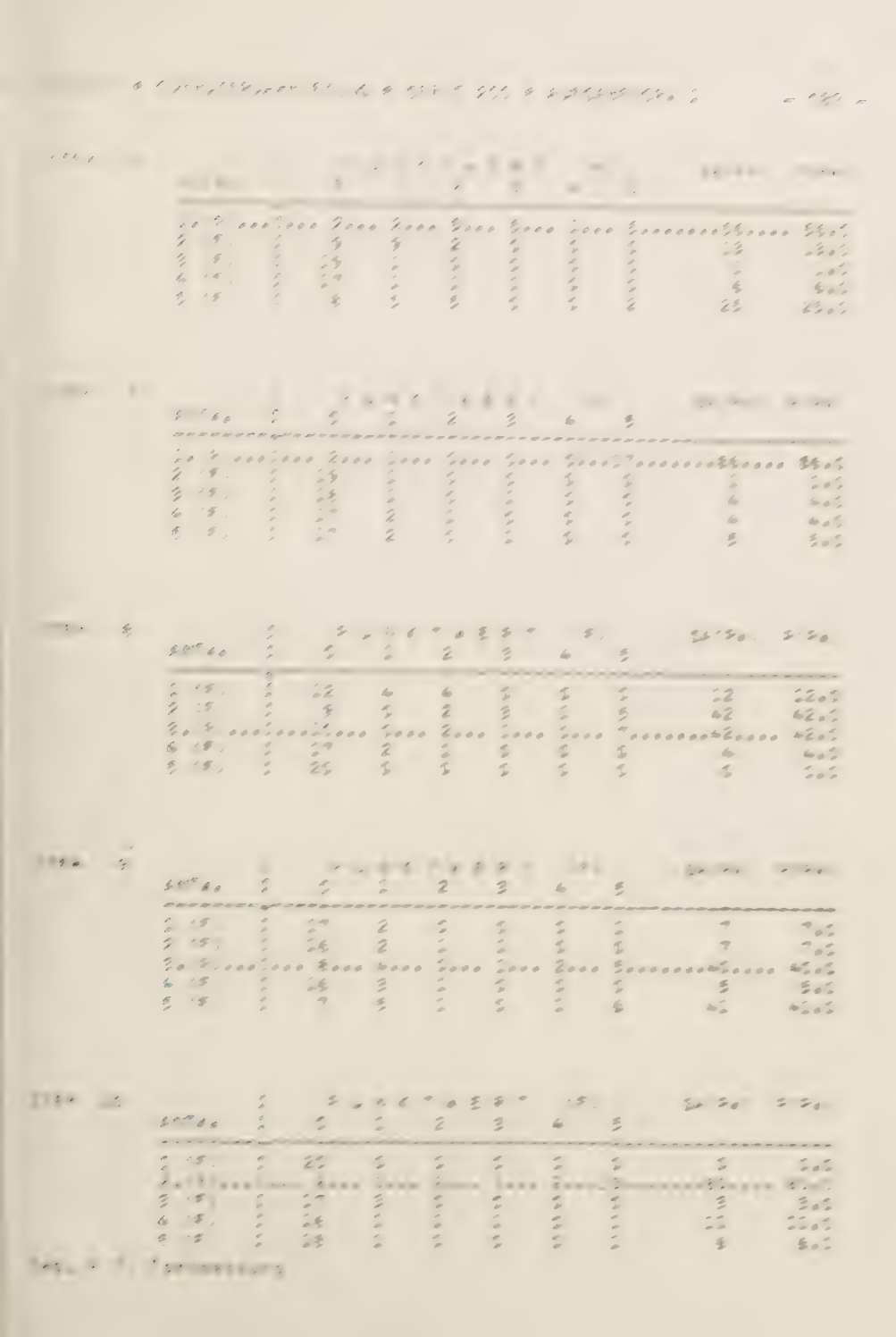
ITEM 4

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	2	0	0	0	0	2	2.0
2 (F)	I	17	2	1	0	0	0	4	4.0
3 (F)	I	18	2	0	0	0	0	2	2.0
4.(R)	I	0	2	0	3	2	13	64	64.0
5 (F)	I	14	4	2	0	0	0	8	8.0

ITEM 5

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	15	3	1	0	1	0	9	9.0
2 (F)	I	16	4	0	0	0	0	4	4.0
3.(R)	I	2	1	3	0	4	10	73	73.0
4 (F)	I	15	3	0	0	0	2	13	13.0
5 (F)	I	19	1	0	0	0	0	1	1.0

Tab. A 7: Häufigkeiten der Punktwerte (F), Summen (SA) und prozentualer Anteil (P) für die 5 Antwortmöglichkeiten der Items des Untertests SATZERGANZUNG in der Versuchsgruppe (2).



ITEM 11		PUNKTWERT (F)						SA(PW)	P(PW)
ANTW.	I	0	1	2	3	4	5		
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	19	1	0	0	0	0	1	1.0
3 (F)	I	19	1	0	0	0	0	1	1.0
4.(R)...	I...	2...	1...	0...	3...	2...	12.....	78....	78.0
5 (F)	I	12	2	4	0	0	2	20	20.0

ITEM 12		PUNKTWERT (F)						SA(PW)	P(PW)
ANTW.	I	0	1	2	3	4	5		
1.(R)...	I...	3...	9...	0...	1...	0...	7.....	47....	47.0
2 (F)	I	9	10	0	0	0	1	15	15.0
3 (F)	I	9	10	0	0	0	1	15	15.0
4 (F)	I	10	9	0	0	0	1	14	14.0
5 (F)	I	11	9	0	0	0	0	9	9.0

ITEM 13		PUNKTWERT (F)						SA(PW)	P(PW)
ANTW.	I	0	1	2	3	4	5		
1 (F)	I	7	9	1	1	0	2	24	24.0
2 (F)	I	11	7	1	0	0	1	14	14.0
3.(R)...	I...	4...	5...	2...	2...	2...	5.....	48....	48.0
4 (F)	I	14	4	2	0	0	0	8	8.0
5 (F)	I	15	4	0	1	0	0	7	7.0

ITEM 14		PUNKTWERT (F)						SA(PW)	P(PW)
ANTW.	I	0	1	2	3	4	5		
1 (F)	I	13	5	1	0	0	1	12	12.0
2 (F)	I	11	7	0	0	0	2	17	17.0
3 (F)	I	12	5	2	0	1	0	13	13.0
4 (F)	I	9	6	2	2	1	0	20	20.0
5.(R)...	I...	4...	8...	1...	3...	1...	3.....	38....	38.0

ITEM 15		PUNKTWERT (F)						SA(PW)	P(PW)
ANTW.	I	0	1	2	3	4	5		
1.(R)...	I...	9...	3...	0...	3...	2...	3.....	35....	35.0
2 (F)	I	9	5	2	1	1	2	26	26.0
3 (F)	I	10	5	0	3	0	2	24	24.0
4 (F)	I	15	2	3	0	0	0	8	8.0
5 (F)	I	17	2	0	0	0	1	7	7.0

Tab. A 7: Fortsetzung

ITEM 16

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1.(R)	I	1	8	1	4	1	5	51	51.0
2 (F)	I	14	5	0	1	0	0	8	8.0
3 (F)	I	13	5	2	0	0	0	9	9.0
4 (F)	I	12	7	0	1	0	0	10	10.0
5 (F)	I	9	5	3	2	0	1	22	22.0

ITEM 17

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	13	5	0	1	0	1	13	13.0
2.(R)	I	5	3	4	2	0	6	47	47.0
3 (F)	I	15	5	0	0	0	0	5	5.0
4 (F)	I	14	4	0	0	1	1	13	13.0
5 (F)	I	10	5	2	1	0	2	22	22.0

ITEM 18

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	19	1	0	0	0	0	1	1.0
2.(R)	I	6	1	3	4	0	6	49	49.0
3 (F)	I	19	0	0	0	0	1	5	5.0
4 (F)	I	9	2	5	2	0	2	28	28.0
5 (F)	I	13	2	2	2	0	1	17	17.0

ITEM 19

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	17	3	0	0	0	0	3	3.0
2 (F)	I	12	6	1	0	0	1	13	13.0
3 (F)	I	17	2	1	0	0	0	4	4.0
4.(R)	I	1	1	3	6	2	7	68	68.0
5 (F)	I	12	5	2	1	0	0	12	12.0

ITEM 20

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	9	7	4	0	0	0	15	15.0
2.(R)	I	12	5	0	2	0	1	16	16.0
3 (F)	I	15	4	0	0	0	1	9	9.0
4 (F)	I	9	9	2	0	0	1	18	18.0
5 (F)	I	4	3	7	2	1	3	42	42.0

Tab. A 7: Fortsetzung

ITEM 1

ANTW.	I	PUNKTWERT 0	1	2	3	4	5 (F)	SA(PW)	P(PW)
1.(R)	I	1	2	0	0	0	17	87	87.0
2 (F)	I	18	2	0	0	0	0	2	2.0
3 (F)	I	17	2	0	0	0	1	7	7.0
4 (F)	I	18	2	0	0	0	0	2	2.0
5 (F)	I	18	2	0	0	0	0	2	2.0

ITEM 2

ANTW.	I	PUNKTWERT 0	1	2	3	4	5 (F)	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	18	0	1	1	0	0	5	5.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4.(R)	I	0	0	1	1	0	18	95	95.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 3

ANTW.	I	PUNKTWERT 0	1	2	3	4	5 (F)	SA(PW)	P(PW)
1 (F)	I	16	0	0	0	0	4	20	20.0
2 (F)	I	19	0	0	0	0	1	5	5.0
3 (F)	I	15	0	0	2	0	3	21	21.0
4.(R)	I	10	0	2	0	0	8	44	44.0
5 (F)	I	18	0	0	0	0	2	10	10.0

ITEM 4

ANTW.	I	PUNKTWERT 0	1	2	3	4	5 (F)	SA(PW)	P(PW)
1 (F)	I	19	0	0	0	0	1	5	5.0
2 (F)	I	18	0	0	0	0	2	10	10.0
3.(R)	I	3	0	0	0	0	17	85	85.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 5

ANTW.	I	PUNKTWERT 0	1	2	3	4	5 (F)	SA(PW)	P(PW)
1.(R)	I	4	0	0	0	0	16	80	80.0
2 (F)	I	16	1	0	0	0	3	16	16.0
3 (F)	I	19	1	0	0	0	0	1	1.0
4 (F)	I	19	0	0	1	0	0	3	3.0
5 (F)	I	20	0	0	0	0	0	0	0.0

Tab. A 8: Häufigkeiten der Punktwerte(F), Summen (SA) und prozentualer Anteil (P) für die 5 Antwortmöglichkeiten der Items des Untertests WORTAUSWAHL in der Versuchsgruppe (2

ITEM 6

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	19	0	1	0	0	0	2	2.0
4 (F)	I	19	1	0	0	0	0	1	1.0
5.(R)	I	0	0	1	0	0	19	97	97.0

ITEM 7

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	19	0	0	0	0	1	5	5.0
4 (F)	I	19	0	0	0	0	1	5	5.0
5.(R)	I	2	0	0	0	0	18	90	90.0

ITEM 8

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	15	0	0	0	0	5	25	25.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4.(R)	I	5	0	0	1	0	14	73	73.0
5 (F)	I	19	0	1	0	0	0	2	2.0

ITEM 9

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3.(R)	I	4	0	1	0	0	15	77	77.0
4 (F)	I	18	0	1	0	0	1	7	7.0
5 (F)	I	16	1	0	0	0	3	16	16.0

ITEM 10

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	16	3	0	1	0	0	6	6.0
2.(R)	I	7	3	1	0	0	9	50	50.0
3 (F)	I	14	3	0	0	0	3	18	18.0
4 (F)	I	15	3	0	0	0	2	13	13.0
5 (F)	I	15	3	0	0	0	2	13	13.0

Tab. A 8: Fortsetzung

ITEM 11

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	1	0	0	0	1	6	6.0
2 (F)	I	19	1	0	0	0	0	1	1.0
3 (F)	I	9	1	0	1	0	9	49	49.0
4.(R)	I	10	1	1	0	0	8	43	43.0
5 (F)	I	19	1	0	0	0	0	1	1.0

ITEM 12

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	16	0	0	0	0	4	20	20.0
3.(R)	I	12	0	0	0	0	8	40	40.0
4 (F)	I	12	0	0	0	0	8	40	40.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 13

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	13	0	1	0	0	6	32	32.0
3 (F)	I	19	0	0	1	0	0	3	3.0
4 (F)	I	18	0	0	0	0	2	10	10.0
5.(R)	I	9	0	0	0	0	11	55	55.0

ITEM 14

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2.(R)	I	9	1	0	0	0	10	51	51.0
3 (F)	I	10	0	0	0	1	9	49	49.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 15

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	14	1	0	1	1	3	23	23.0
4.(R)	I	3	1	1	0	1	14	77	77.0
5 (F)	I	20	0	0	0	0	0	0	0.0

Tab. A 8: Fortsetzung

ITEM 16

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	1	1	0	0	0	3	3.0
2.(R)	I	4	1	0	0	0	15	76	76.0
3 (F)	I	18	1	1	0	0	0	3	3.0
4 (F)	I	18	1	0	1	0	0	4	4.0
5 (F)	I	16	1	0	1	0	2	14	14.0

ITEM 17

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1.(R)	I	14	0	0	0	0	6	30	30.0
2 (F)	I	19	0	0	0	0	1	5	5.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	14	0	0	0	0	6	30	30.0
5 (F)	I	13	0	0	0	0	7	35	35.0

ITEM 18

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	19	0	1	0	0	0	2	2.0
2 (F)	I	19	0	0	0	0	1	5	5.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4.(R)	I	7	0	1	0	0	12	62	62.0
5 (F)	I	13	1	0	0	0	6	31	31.0

ITEM 19

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	17	1	0	0	0	2	11	11.0
2.(R)	I	9	2	0	0	0	9	47	47.0
3 (F)	I	18	1	0	1	0	0	4	4.0
4 (F)	I	13	1	1	0	0	5	28	28.0
5 (F)	I	17	1	0	0	1	1	10	10.0

ITEM 20

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	4	1	0	0	1	14	75	75.0
2 (F)	I	19	1	0	0	0	0	1	1.0
3.(R)	I	14	1	0	1	0	4	24	24.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

Tab. A 8: Fortsetzung

ITEM 1

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1.(R)	I	0	0	0	0	0	20	100	100.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 2

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	19	0	0	0	0	1	5	5.0
2.(R)	I	3	0	0	0	0	17	85	85.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	18	0	0	0	0	2	10	10.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 3

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	19	0	0	1	0	0	3	3.0
3.(R)	I	1	0	1	0	0	18	92	92.0
4 (F)	I	19	0	0	0	0	1	5	5.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 4

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	19	0	0	0	0	1	5	5.0
3 (F)	I	19	0	0	0	0	1	5	5.0
4.(R)	I	3	0	0	0	0	17	85	85.0
5 (F)	I	19	0	0	0	0	1	5	5.0

ITEM 5

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	19	0	0	1	0	0	3	3.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3.(R)	I	1	0	1	1	2	15	88	88.0
4 (F)	I	16	2	1	1	0	0	7	7.0
5 (F)	I	19	0	1	0	0	0	2	2.0

Tab. A 9: Häufigkeiten der Punktwerte (F), Summen (SA) und prozentualer Anteil (P) für die 5 Antwortmöglichkeiten der Items des Untertests ANALOGIEN in der Versuchsgruppe (2)

ITEM 6

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	0	1	0	0	1	7	7.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5.(R)	I	1	0	0	1	0	18	93	93.0

ITEM 7

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	20	0	0	0	0	0	0	0.0
2 (F)	I	19	0	1	0	0	0	2	2.0
3 (F)	I	19	0	0	0	0	1	5	5.0
4 (F)	I	19	0	0	1	0	0	3	3.0
5.(R)	I	2	0	0	0	0	18	90	90.0

ITEM 8

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	16	0	1	1	0	2	15	15.0
2.(R)	I	7	0	2	1	0	10	57	57.0
3 (F)	I	18	1	1	0	0	0	3	3.0
4 (F)	I	14	0	4	0	0	2	18	18.0
5 (F)	I	17	2	0	0	0	1	7	7.0

ITEM 9

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	17	1	0	1	0	1	9	9.0
2 (F)	I	19	1	0	0	0	0	1	1.0
3 (F)	I	19	1	0	0	0	0	1	1.0
4.(R)	I	3	1	1	0	0	15	78	78.0
5 (F)	I	17	1	0	0	0	2	11	11.0

ITEM 10

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	1	0	1	0	0	4	4.0
2 (F)	I	15	2	1	0	0	2	14	14.0
3.(R)	I	4	2	1	2	1	10	64	64.0
4 (F)	I	13	4	2	0	0	1	13	13.0
5 (F)	I	15	5	0	0	0	0	5	5.0

Tab. A 9: Fortsetzung

ITEM 11

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	14	1	0	1	0	4	24	24.0
2.(R)	I	4	0	1	0	1	14	76	76.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 12

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	19	0	0	0	0	1	5	5.0
2 (F)	I	18	0	0	1	0	1	8	8.0
3.(R)	I	5	0	1	0	0	14	72	72.0
4 (F)	I	19	0	0	1	0	0	3	3.0
5 (F)	I	17	0	1	0	0	2	12	12.0

ITEM 13

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1.(R)	I	3	1	0	0	0	16	81	81.0
2 (F)	I	18	2	0	0	0	0	2	2.0
3 (F)	I	17	1	0	1	0	1	9	9.0
4 (F)	I	17	2	0	0	0	1	7	7.0
5 (F)	I	19	1	0	0	0	0	1	1.0

ITEM 14

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1.(R)	I	5	2	0	0	1	12	66	66.0
2 (F)	I	18	2	0	0	0	0	2	2.0
3 (F)	I	14	5	0	0	0	1	10	10.0
4 (F)	I	18	2	0	0	0	0	2	2.0
5 (F)	I	14	2	0	0	2	2	20	20.0

ITEM 15

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	16	1	1	0	0	2	13	13.0
2 (F)	I	19	1	0	0	0	0	1	1.0
3 (F)	I	15	2	0	0	0	3	17	17.0
4 (F)	I	17	2	1	0	0	0	4	4.0
5.(R)	I	5	1	2	0	0	12	65	65.0

Tab. A 9: Fortsetzung

ITEM 16

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	17	0	0	0	0	3	15	15.0
2 (F)	I	11	0	1	0	1	7	41	41.0
3.(R)	I	13	0	0	0	0	7	35	35.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	17	1	0	1	0	1	9	9.0

ITEM 17

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	10	0	0	0	0	10	50	50.0
2.(R)	I	10	0	0	0	0	10	50	50.0
3 (F)	I	20	0	0	0	0	0	0	0.0
4 (F)	I	20	0	0	0	0	0	0	0.0
5 (F)	I	20	0	0	0	0	0	0	0.0

ITEM 18

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	17	0	1	0	0	2	12	12.0
2 (F)	I	19	0	1	0	0	0	2	2.0
3 (F)	I	19	1	0	0	0	0	1	1.0
4.(R)	I	4	0	0	0	0	16	80	80.0
5 (F)	I	18	0	1	1	0	0	5	5.0

ITEM 19

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	18	1	1	0	0	0	3	3.0
2.(R)	I	13	1	0	0	0	6	31	31.0
3 (F)	I	6	1	0	1	0	12	64	64.0
4 (F)	I	19	1	0	0	0	0	1	1.0
5 (F)	I	19	1	0	0	0	0	1	1.0

ITEM 20

ANTW.	I	PUNKTWERT (F) 0	1	2	3	4	5	SA(PW)	P(PW)
1 (F)	I	8	0	3	2	0	7	47	47.0
2 (F)	I	20	0	0	0	0	0	0	0.0
3.(R)	I	13	0	2	0	0	5	29	29.0
4 (F)	I	13	1	2	1	0	3	23	23.0
5 (F)	I	19	1	0	0	0	0	1	1.0

Tab. A 9: Fortsetzung

PROGRAMMANHANG

Der Programmanhang enthält die Auflistung folgender FORTRAN-Unterprogramme für das automatisierte Testsystem:

ADRUP, DAEUP, DGBUP, DIGUP, DTXUP, DZFUP, FELUP, PAGUP, PENUP, PRNUP, STRUP, TASUP, TSTUP, ZFZUP, ZTVGL.

```
      SUBROUTINE ADRUP (NTEX, IAN, IAUS,ADDR)
C
      INTEGER ADDR(1)
      IAN = ADDR(NTEX)
      IAUS = ADDR(NTEX + 1) - 1
C
      RETURN
      END




      SUBROUTINE DAEUP (TA1,IT1,TA2,IT2,ITRC,     IZAHL,IT,IUW,IOW,
     $                  IMONA,NMON,IRZMS)
C
      INTEGER TA1(1), TA2(1), WERT(4)
      COMMON KOR, IZZZZ, RZZZZ
      IZM = IRZMS
      RZZZZ = 0.
      KOR = 0
C
    1 CONTINUE
C
C   STANDARD- ODER ZUSATZTEXTFILE
C
      GO TO (10,20), IT
C
C   ADDRESSENBERECHNUNG FUER STANDARDTEXTFILE
C
   10 CONTINUE
      NTXRC = 3101
      CALL ADRUP (IT1,IAN,IAUS,TA1)
      GO TO 30
C
C   ADDRESSENBERECHNUNG FUER ZUSATZTEXTFILE
C
   20 CONTINUE
      NTXRC = ITRC
      CALL ADRUP (IT2,IAN,IAUS,TA2)
C
C   BILDSCHIRMAUSGABE DES TEXTES
C
   30 CONTINUE
      CALL DTXUP (IAN,IAUS,NTXRC)
C
C     BILDSCHIRMAUSGABE '=ANTWORT='
C
      WRITE (4,2000)
 2000 FORMAT (T66,'=== ANTWORT ==='/T69,1X)
C
C   ABFRAGE DER EINGABETASTATUR
C
   35 CONTINUE
      IWERT = 1
      CALL SEZEL (WERT,4,1,0)
C
   40 CONTINUE
      CALL DZEUP (IWERT,IP,IRTUM)
C
C   MESSUNG DER ZEIT BIS ZUR ERSTEN REAKTION
```

```
      IF (IZM)                    45,45,42
 42   CONTINUE
      CALL ITIME
      RZZZZ = IZZZZ
      RZZZZ = RZZZZ / 100.
      IZM = 0
 45   CONTINUE
C
C    WEITERVERARBEITUNG DER EINGETASTETEN INFORMATION
C
      GO TO (50,50,60,70,5,80), IRTUM
C
C     EINGABE KORREKTUR
C
    5 CONTINUE
      KOR = KOR  + 1
      WRITE (4,2001)
 2001 FORMAT ('+ ', T70, 5(' '))
      GO TO 35
C
C    FEHLERHAFTE EINGABE
C
C    UNERWARTETE INFORMATION
C
   50 CONTINUE
      ITX = 8
      GO TO 65
C
C    ZAHL UMFASST MEHR ALS VIER STELLEN
C
   60 CONTINUE
      ITX = 9
C
C    ANZEIGE DES FEHLERS AUF DEM BILDSCHIRM
C
   65 CONTINUE
      CALL STRUP (TA1,ID,ID,ITX,1)
      GO TO 1
C
C    RICHTIGE EINGABE
C    BILDSCHIRMAUSGABE DER EINGETASTETEN ZIFFER
C
   80 CONTINUE
      CALL PENUP (IP,IWERT,WERT,IZAHL)
      GO TO 40
C
C     EINGABE ENDE - RICHTIG
C
   70 CONTINUE
      IWERT = IWERT - 1
      CALL ZFZUP (WERT,IWERT,IZAHL)
C
C     FEHLERANZEIGE BEI GRENZUEBERSCHREITUNG DER EINGEGEBN ZAHL
C
      IF (IFRLT(IZAHL,IUW,IOW))100,99,100
 100  CONTINUE
      CALL STRUP (TA1,TA2,ITRC,NMON,IMONA)
      GO TO                       1
C
 99   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE DGRUP (IUW, IOW, IP, IRTUM)
C
      CALL IFDIK (IP)
      IA=-10
      IP = 0
    1 CONTINUE
C
      IF (IFDIK(IP)) 1,1,2
    2 CONTINUE
    3 CONTINUE
      IF(IFDIK(IA)) 3,4,3
    4 CONTINUE
C
      IRTUM = 2 -IABS (IFRLT(IP,IUW,IOW))
C
      RETURN
      END




      SUBROUTINE  DIGUP (IW)
C
      IP = -1
    1 CONTINUE
      IF (IFDIK(IP) - IW)    1,2,1
    2 CONTINUE
    3 CONTINUE
      IF(IFDIK(IP)) 3,4,3
    4 CONTINUE
      RETURN
      END




      SUBROUTINE DTXUP (IAN,IAUS,ITXRC)
      INTEGER KARTE (40), ZEILE(39), VOR
      EQUIVALENCE (KARTE(1),VOR) , (KARTE(2), ZEILE(1))
 2000 FORMAT ('+',39A2)
      DO 11 I = IAN, IAUS
      READ (ITXRC'I) KARTE
      CALL PAGUP (VOR)
      WRITE (4,2000) ZEILE
   11 CONTINUE
      RETURN
      END
```

```
      SUBROUTINE DZEUP (IWERT, IP, IRTUM)
C
      CALL DGBUP (1,10,IP,IRTUM)
C
      GO TO (10,20), IRTUM
C
C   FUNKTIONSTASTE GEDRUECKT
C
   10 IF (IWERT - 1) 30,30,40
C
C   FUNKTIONSTASTE VOR ZIFFERN (FEHLER)
C
   30 IRTUM = 1
      RETURN
C
C
C   ABFRAGE OB DER EINGABE WERT 15 UEBERSTEIGT
C
   40 II = IP - 10
      IF (IFRLT(II,1,5)) 901,101,901
  901 CONTINUE
      II = 5
  101 CONTINUE
C
      GO TO (50,60,70,50,50), II
C   START, 14 ODER 15
   50 IRTUM = 2
      RETURN
C
C   EINAGBE ENDE (RICHTIG)
C
   60 IRTUM = 4
      RETURN
C
C   EINGABE KORREKTUR (RICHTIG)
C
   70 IRTUM = 5
      RETURN
C
C   ZAHLENEINGABE
C
   20 IF (4-IWERT) 80,90,90
C
C   MEHR ALS VIER STELLEN (FEHLER)
C
   80 IRTUM = 3
      RETURN
C
C   BIS ZU VIER STELLEN (RICHTIG)
C
   90 IRTUM = 6
C
      RETURN
      END
```

```
      SUBROUTINE FELUP (ERG,RICHT,WERT,NITEM,IPERM,NRICH)
      INTEGER ERG(1), RICHT(1), WERT(1)
      NRICH = 0
C
      IF (IPERM - 2)            10,100,10
 10   CONTINUE
      DO 11 I = 1,NITEM
      IF(ERG(I) - RICHT(I))     15,20,15
 15   CONTINUE
      WERT(I) = 0
      GO TO                     11
 20   CONTINUE
      WERT(I) = 1
      NRICH = NRICH + 1
 11   CONTINUE
      RETURN
C
 100  CONTINUE
      DO 22 I = 1,NITEM
      IR= RICHT(I)
      IZ = IR / 10
      IZ = IZ * 10
      IE = IR - IZ
      IZ = IZ/10 + IE * 10
      IF (ERG(I).EQ.IR.OR.ERG(I).EQ.IZ) GO TO 30
      WERT(I) = 0
      GO TO 22
 30   CONTINUE
      WERT(I) = 1
      NRICH = NRICH + 1
 22   CONTINUE
      RETURN
      END




      SUBROUTINE PENUP (IP, IWERT, WERT, IZAHL)
C
      INTEGER WERT(4)
      WERT (IWERT) = IP - IP/10 * IP
      CALL ZFZUP (WERT,IWERT,IZAHL)
      WRITE (4,2000) IZAHL
2000  FORMAT ('+',T70,I5,T69,1X)
      IWERT = IWERT + 1
      RETURN
      END
```

```
      SUBROUTINE PAGUP(IZ)
      IF (IZ)                    10,30,20
 10   CONTINUE
      WRITE (4,2000)
 2000 FORMAT ('1 ')
      IZ = IABS(IZ) - 1
      IF (IZ)                    30,30,20
 20   CONTINUE
      DO 11 I = 1,IZ
      WRITE (4,2001)
 2001 FORMAT (' ')
 11   CONTINUE
 30   RETURN
      END



      SUBROUTINE PRNUP (DD,NVP,ERG,WERT,RZT,ZEIT,KR,BEARB,NRICH,IGZ,IST)
      REAL ZEIT (1), RZT(1)
      INTEGER DD(4), ERG(1), KR(1), BEARB(1), WERT(1), NRICH(1), IGZ(1)
 2000 FORMAT ('1',T2,80('=')/T10,'HAUPTVERSUCH G. GLAUBITZ',T65,4A2/
     *T2,80('=')///T10,'VERSUCHSPERSON NR.'I8,T55,'AUFGABENGRUPPE',
     *I4///)
 2001 FORMAT (
     19X,'ITEM      ANTW.  WERT      R-ZEIT  G-ZEIT      N-KORR  N-BEARB
     2 ' / T2, 40('=-') // )
 2002 FORMAT ('0',T12,I2,T20,2I5,T36,2F8.2,T54,2I8)
 2003 FORMAT ('0',T28,4('-'),T46,6('-')/T20,'SUMME',I5,T35,'TESTZEIT',
     1 I6,'.00')
 3000 FORMAT (32('*'),4A2,32('*'),'GH',I6)
 3001 FORMAT(2I5,2F10.2,2I5,T73,'GH',2I3)
 3002 FORMAT (20I1,2I5,T73,'GH',2I3)
C     AUSSCHRIEBEN DER ERGEBNISSE
      DO 33                      IAUF = 1,3
      WRITE (3,2000) DD,NVP,IAUF
      WRITE (3,2001)
      IX = 20 * IAUF
      IA = IX - 19
      DO 44                      J = IA,IX
      WRITE (3,2002) J,ERG(J),WERT(J),RZT(J),ZEIT(J),KR(J),BEARB(J)
 44   CONTINUE
      WRITE (3,2003) NRICH(IAUF), IGZ(IAUF)
 33   CONTINUE
      IF (IST)                   90,90,5
 5    CONTINUE
C     STANZEN DER ERGEBNISSE
      CALL FKORL
      WRITE (2,3000) DD, NVP
      DO 55                      I = 1,60
      CALL FKORL
      WRITE (2,3001)  ERG(I),WERT(I),RZT(I),ZEIT(I),KR(I),BEARB(I),I,NVP
 55   CONTINUE
      DO 66                      I = 1,3
      ILAB = 100 + I
      CALL FKORL
      IX = I * 20
      IA = IX - 19
      WRITE (2,3002) (WERT(J),J=IA,IX), NRICH(I),IGZ(I),ILAB,NVP
 66   CONTINUE
 90   CONTINUE
      RETURN
      END
```

```
      SUBROUTINE STRUP (STXAD,TXAD,ITXRC,NTX,IT)
      INTEGER STXAD(1), TXAD(1)
      GO TO                     (10,20), IT
10    CONTINUE
      NTXRC = 3101
      CALL ADRUP (NTX,IAN,IAUS,STXAD)
      GO TO                     30
20    CONTINUE
      NTXRC = ITXRC
      CALL ADRUP (NTX,IAN,IAUS,TXAD)
30    CONTINUE
      CALL DTXUP (IAN,IAUS,NTXRC)
      WRITE (4,2000)
2000  FORMAT (T74,'/START/')
      CALL DIGUP(11)
      RETURN
      END
```

```
      SUBROUTINE TASUP (STXAD)
      INTEGER STXAD(1), IUN(5), IOB(5), MON(5)
      DATA IUN/0,0,1,120,1/ , IOB/9999,9999,3,240,2/ , MON/9,9,12,10,12/
      CALL DAFUP (STXAD,11,ID,ID,ID,IZAHL,1,1,2,1,12,0)
      IF (IZAHL - 1)            99,10,99
10    CONTINUE
      DO 11 I = 1,2
      CALL STRUP (STXAD,ID,ID,I,1)
11    CONTINUE
      DO 22 I = 1,5
      II = 2 + I
      CALL DAFUP (STXAD,II,ID,ID,ID,IZAHL,1,IUN(I),IOB(I),1,MON(I),0)
22    CONTINUE
      IF (IZAHL - 1)            10,99,10
99    CONTINUE
      RETURN
      END
```

```
      SUBROUTINE TSTUP (STXAD,TXAD,ITXRC,NITEM,ITEMA,IZT,NWIE,ZEIT,ERG,
     $ KR,BEARB,INRPT,IU,IO,RZT,IGZ)
      REAL ZEIT(1), RZT(1)
      INTEGER STXAD(1),TXAD(1),ERG(1),KR(1),BEARB(1)
      COMMON KOR, IZZZZ, RZZZZ
C
      IZEIT = IZT
      IDUR = 0
      INRPT = 0
      ASSIGN 900 TO IADR
      CALL RZEIT (IZEIT,IADR,0)
C
      DO 11                         I = 1,NITEM
      RZT (I) = 0.
      ZEIT(I) = 0.
      ERG(I) = 0
      KR(I) = 0
      BEARB(I) = 0
 11   CONTINUE
C
 10   CONTINUE
      K = 0
      ITEM = ITEMA - 1
      DO 22                         I = 1,NITEM
      ITEM = ITEM + 1
      K = K + 1
      IF (ERG(K))                   15,15,22
 15   CONTINUE
      CALL IINT
      CALL DAEUP (STXAD,ID,TXAD,ITEM,ITXRC,IZAHL,2,IU,IO ,1,12,1)
      CALL IINT
      EZ = IZZZZ
      EZ = EZ / 100.
      IF (IDUR)                     12,12,14
 12   CONTINUE
      RZT (K) = RZZZZ
 14   CONTINUE
      IF (IZEIT)                    900,900,16
 16   CONTINUE
      ZEIT(K) = ZEIT(K) + EZ
      KR(K) = KR(K) + KOR
      BEARB(K) = BEARB(K) + 1
      IF (NWIE)                     20,20,17
 17   CONTINUE
      IF (IZAHL-9999)               20,22,20
 20   CONTINUE
      ERG(K) = IZAHL
 22   CONTINUE
      IDUR = 1
C
      IF (NWIE)                     999,999,30
 30   CONTINUE
      DO 33 M = 1,NITEM
      IF (ERG(M))                   10,10,33
 33   CONTINUE
      GO TO                         999
C
 900  CONTINUE
      ZEIT (K) = ZEIT(K) + EZ
      KR(K) = KR(K) + KOR
      BEARB(K) = BEARB(K) + 1
      INRPT = 1
C
 999  CONTINUE
```

```
      IGZ = IABS (IZT - IZEIT)
      CALL RZAUS
      RETURN
      END




      SUBROUTINE ZFZUP (IZIF,ISTEL, IZAHL)
C
      DIMENSION IZIF(1)
      IZAHL = 0
C
      DO 11 I = 1, ISTEL
      J = ISTEL - I
   11 IZAHL = IZAHL + 10 ** J * IZIF(I)
      RETURN
      END




      SUBROUTINE ZTVGL
 1    CONTINUE
      IZT = 20
      ASSIGN 900 TO IA
      CALL RZEIT (IZT,IA,0)
      CALL TIME (8,IZZ)
      CALL IINT
 5    CONTINUE
      CALL ITIME
      IF (1000-IZZ)            10,10,5
 10   CONTINUE
      CALL TIME (0,IZZ)
 900  CONTINUE
      K = 20 - IZT
      CALL RZAUS
      IF (K.GE.9.AND.K.LE.11) GO TO 99
      CALL PAGUP (-9)
      WRITE (4,4000)
4000  FORMAT (T20,'*** KONTROLLE ZEITSCHALTER ***')
      WRITE (4,4100) K
4100  FORMAT (I10)
      PAUSE 1111
      GO TO                    1
99    RETURN
      END
```

LEBENSLAUF

Am 31. März 1942 wurde ich als Sohn des Dipl.-Landwirtes Dr. Wilhelm Plessen und seiner Ehefrau Dipl.-Volkswirtin Christa Plessen, geb. Ruhnke, in Berlin geboren. Nach dem Besuch der Volksschule in Mühlenbeck/Berlin und ab 1950 in Köln-Nippes besuchte ich von 1952 bis 1958 das Neusprachliche Gymnasium mit altsprachlichem Zweig in Köln-Nippes und ab Ostern 1958 das altsprachliche Max-Planck-Gymnasium in Göttingen, an dem ich im Februar 1963 die Reifeprüfung ablegte.

Im Sommersemester 1963 begann ich mein Studium mit dem Hauptfach Germanistik an der Universität Innsbruck, studierte im Wintersemester 1963/64 Germanistik und Psychologie an der Universität Göttingen bzw. im Sommersemester 1964 an der Technischen Hochschule Hannover und nahm mit dem Wintersemester 1964/65 das eigentliche Studium der Psychologie an der Universität Münster auf. Hier legte ich im November 1968 die Diplom-Vorprüfung und im Februar 1970 die Diplom-Hauptprüfung für Psychologen ab.

Von April 1969 bis Februar 1970 bin ich als studentische Hilfskraft, von April 1970 bis Oktober 1971 als wissenschaftliche Hilfskraft in der Klinischen Abteilung des Psychologischen Institutes der Universität Münster tätig gewesen. Bis zum Oktober 1973 erhielt ich ein Stipendium nach dem Graduierten-Förderungs-Gesetz und hatte ab Sommersemester 1972 im Fachbereich Psychologie der Universität Münster einen Lehrauftrag für das Gebiet der Psychologischen Diagnostik. Seit Oktober 1973 bin ich als Verwalter einer Wissenschaftlichen Assistentenstelle in der Klinischen Abteilung des Psychologischen Institutes der Universität Münster in Forschung und Lehre mit dem Schwerpunkt "Klinisch-Psychologische Diagnostik und Beratung" tätig.

Von meinen akademischen Lehrern bin ich vor allem Frau Prof. Dr. Lilly Kemmler, Herrn Prof. Dr. H. Bommert, Herrn Prof. Dr. Dr. h.c. W. Metzger und Herrn Prof. Dr. W. Witte wegen ihres nachhaltigen Einflusses auf meinen wissenschaftlichen Werdegang zu besonderem Dank verpflichtet.

B2159.09